GENOMIC INSIGHTS INTO *MYCOBACTERIUM TUBERCULOSIS* AND ITS INTERACTION WITH THE MICROBIOTA

by
Kathryn Winglee

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
May 2015

Abstract

Tuberculosis (TB) is the leading cause of death from a bacterial infection in humans. Despite its impact throughout history on humans across the globe, it remains challenging to diagnose and treat. This work used molecular biology and next generation sequencing to explore these issues. First, in a study to identify potential biomarkers of TB infection, the interaction between *Mycobacterium tuberculosis* (the causative agent of TB), the mouse immune system, and the murine gut microbiota was examined. The murine gut microbiota was observed to respond specifically to *M. tuberculosis* infection in several host genotypes, and these changes were most likely mediated by the adaptive immune system. Together, these data confirm that the response of the gut microbiota can be further explored for TB diagnostics. A second study was aimed at understanding the genetic mechanisms of resistance to a novel anti-mycobacterial compound. Resistance was mediated through loss of function of *Rv2887,* a previously unannotated gene that was found to be a multiple antibiotic resistance repressor (MarR) transcriptional regulator. Analysis of the function of *Rv2887* led to the identification of a gene regulation mechanism that could be a potential new drug target. Finally, in a third study the genetic basis of geographic restriction of *M. africanum*, a mycobacterial species that causes similar disease to human TB but is usually only found in West Africa, was elucidated. Despite conventional dogma, analysis of *M. africanum* using new bioinformatics tools revealed that it is not a separate species from *M. tuberculosis*. Furthermore, *M. africanum* is unimpaired in transmission or virulence compared to *M. tuberculosis*, thus suggesting that the geographic restriction may be due to host factors. Taken together, this work explores the host-pathogen interactions and genetics of mycobacteria and provides novel insights into how these bacteria cause TB.

Thesis Advisor:

William R. Bishai, MD, PhD, Professor of Medicine and Pathology, Johns Hopkins University School of Medicine

Second Reader:

Srinivasan Yegnasubramanian, M.D., Ph.D., Assistant Professor of Oncology, Johns Hopkins University School of Medicine

Thesis Committee:

Srinivasan Yegnasubramanian, M.D., Ph.D., Assistant Professor of Oncology, Johns Hopkins University School of Medicine (Chair)

Brendan Cormack, Ph.D., Professor of Molecular Biology and Genetics, Johns Hopkins University School of Medicine

Joel Bader, Ph.D., Associate Professor of Biomedical Engineering, Johns Hopkins University School of Medicine

Acknowledgments

My Ph.D. was a great learning experience, and there are a number of people that made it possible. First, I would like to thank my Ph.D. supervisor, Dr. William Bishai, who gave me numerous opportunities to explore different topics and helped to foster collaborations that greatly enriched this experience. I would also like to thank my thesis committee, Dr. Srinivasan Yegnasubramanian, Dr. Brendan Cormack and Dr. Joel Bader for their valuable time and insight, especially Dr. Yegnasubramanian for chairing my committee and reading my thesis.

I would also like to acknowledge all past and present members of the Center for Tuberculosis Research, and particularly the members of the Bishai lab. All of you have made this a wonderful work environment and have taught me a lot. In particular, I would like to thank Dr. Shashank Gupta, who provided mentorship, advice and friendship. We have had an amazing time working together and none of this would have been possible without him.

I would also like to thank all of my collaborators outside of Johns Hopkins University, particularly Dr. Ashlee Earl at the Broad Institute for providing advice and guidance and giving me the opportunity to work closely with her team.

For their assistance with chapter 2, I would additionally like to acknowledge Dr. Emiley Eloe-Fadrosh and Dr. Claire Fraser from the University of Maryland. For chapter 3, I would like to thank Dr. Curtis Huttenhower and Dr. Galeb Abu-Ali from Harvard School of Public Health for their guidance on data analysis. For chapter 4, I would like to thank Dr. Shichun Lun from Johns Hopkins University, Dr. Marco Pieroni from Università di Parma and Dr. Alan Kozikowski from the University of Illinois at Chicago for their assistance with the compound. For chapter 5, I would like to thank Dr. Thomas Abeel, Dr. Terrance Shea, Dr. Christopher Desjardins, and especially Dr. Abigail Manson McGuire from the Broad Institute for their assistance with

sequencing analysis. In addition, this project would not have been possible without Dr. Mamoudou Maiga, now at the National Institute of Health, and everyone at SEREFO.

Finally, I would like to thank my parents, Jennifer and Robert Winglee, my brother Matthew, and the rest of my family for their support all of these years. I would not have gotten here without them.

There are a number of friends and colleagues who were not mentioned here, but who have still had a positive impact on this work. Thank you to everyone for your support and guidance during the past 6 years.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Mycobacterium tuberculosis is an acid fast bacterium that causes tuberculosis (TB), a devastating disease that affects one-third of the world's population and kills 1.3 million people a year, mostly in developing countries [1]. After HIV, it is the second deadliest infectious agent, and it is a leading killer of HIV-infected patients. *M. tuberculosis* is spread through the aerosol route and primarily causes pulmonary disease, although it can spread throughout the body [1].

Tuberculosis has evolved alongside its human host, and has been identified in ancient Egyptian and Peruvian mummies [2-4]. Despite advances in modern medicine, tuberculosis remains challenging and time consuming to treat. Standard tuberculosis treatment is currently six months long, and involves a two-month high intensity phase consisting of treatment with four antibiotics (isoniazid, rifampin, ethambutol, and pyrazinamide) followed by a four month continuation phase (treat with isoniazid and rifampin only). However, despite this combination therapy, drug resistance is on the rise, and in 2013, there were an estimated 480,000 cases of multi-drug resistant tuberculosis (MDR-TB), which is defined as bacteria that is resistant to at least isoniazid and rifampin. These cases require up to two years of treatment, but even under these conditions drug resistance is developing, and 9% of MDR-TB cases were extremely drug resistant (XDR), meaning that they were also resistant to second line antibiotics [1]. Worse, there are now reports of tuberculosis patients diagnosed with totally drug resistant (TDR) tuberculosis, which are resistant to all tested antibiotics, leaving patients and physicians with few options [5,6].

Not only does long treatment and developing drug resistance provide barriers to eradication of tuberculosis, but diagnosis also remains challenging. *M. tuberculosis* is slow growing, with a doubling time of 24 hours in culture, and thus standard culture based

approaches can take at least a month, even under ideal conditions [7]. Newer approaches, such as the Gene Xpert, use mycobacterial-specific sequences to address these issues [8]. However, these are limited by a lack of knowledge of the genetics of *M. tuberculosis* and what different mutations mean.

One new tool for uncovering genetic mechanisms is next generation sequencing. This provides a relatively cost effect way to analyze the nucleic acid (RNA or DNA) content of a sample. There are several such technologies currently available, including the Roche 454, Illumina HiSeq and MiSeq, ABI SOLiD, PacBio, Ion Torrent and Ion Proton[9,10]. All of these are based on the principle of massively parallel sequencing of millions of nucleic acid fragments and then using computational bioinformatics approaches to interpret the resulting reads to develop meaningful inferences, such as the whole genome sequence of a clinical isolate. Although currently too expensive to be used routinely in diagnostics, the cost of sequencing is rapidly decreasing, making it feasible to ask new questions about the biology of pathogens and their interaction with their host.

In this work, we used a combination of molecular biology techniques and next generation sequencing to analyze the interaction of *M. tuberculosis* with the microbiome, identify the target of a novel anti-mycobacterial compound, and study the genetics of Mali clinical isolates. Together, this data i) explores the interaction between pathogen, host, and resident microbial community, ii) identifies a novel mechanism of gene regulation in mycobacteria and iii) provides new insights into genetics of *M. tuberculosis* and the closely related *M. africanum*.

Chapter 2: Aerosol Mycobacterium tuberculosis Infection Causes Rapid Loss of Diversity in Gut Microbiota

This work has been published in PLoS ONE, doi: 10.1371/journal.pone.0097048, with assistance from Dr. Emiley Eloe-Fadrosh for data analysis and figure generation.

2.1 Abstract

Mycobacterium tuberculosis is an important human pathogen, and yet diagnosis remains challenging. Little research has focused on the impact of *M. tuberculosis* on the gut microbiota, despite the significant immunological and homeostatic functions of the gastrointestinal tract. To determine the effect of *M. tuberculosis* infection on the gut microbiota, we followed mice from *M. tuberculosis* aerosol infection until death, using 16S rRNA sequencing. We saw a rapid change in the gut microbiota in response to infection, with all mice showing a loss and then recovery of microbial community diversity, and found that pre-infection samples clustered separately from post-infection samples, using ecological beta-diversity measures. The effect on the fecal microbiota was observed as rapidly as six days following lung infection. Analysis of additional mice infected by a different *M. tuberculosis* strain corroborated these results, together demonstrating that the mouse gut microbiota significantly changes with *M. tuberculosis* infection.

2.2 Introduction

Although the mouse immunopathologic response to TB differs from the response in humans, the mouse model is frequently used to study the virulence of different strains of *M. tuberculosis*, and to assess TB drug efficacy [11]. Upon aerosol infection, the bacteria promptly replicate in the lungs, and disseminated bacilli are observed in the spleen and liver two to four weeks later. Beginning at four weeks post-infection, the mouse adaptive immune system

achieves partial control of the infection, resulting in a plateau in bacterial burden [12]. The mouse will develop granuloma-like lesions in the lung and, with a high infecting dose, will eventually succumb to the bacteria and die.

One aspect of TB infection that remains largely unexplored is the role of and impact on the resident microbiota (the microorganisms that collectively live on or in mammals). It is now well recognized that the gastrointestinal microbiota maintain a complex, reciprocal relationship with the host immune system [13-16]. Moreover, differences in fecal microbiota composition and functional potential have been identified in individuals with various disease states when compared to healthy individuals, including inflammatory bowel disease, arthritis, type 2 diabetes, and asthma [17-19]. The gut microbiota also plays a prominent role in both enteric bacterial and viral infections [20-23]. However, there have been few studies focused on changes in the gut microbiota in response to TB, especially in the absence of antibiotics.

At the beginning of our research, only one other study had focused on the gut microbiota during TB infection. Dubourg and colleagues utilized culture-dependent and -independent methods to evaluate bacterial and fungal diversity within a single patient at a single time, and found that there was an impoverished community dominated by only a few phylotypes [24]. However, this patient had been treated with broad-spectrum antibiotics for four months, which can greatly alter the gut microbiota. Another study concluded that there was a difference between the sputum microbial composition of TB patients and healthy controls [25]. Conversely, a different paper focused on human sputum samples found no difference between TB patients and healthy controls [26].

Given the interaction between the resident microbiota and the immune system, as well as competition among the microbiota and invading pathogens, we hypothesized that *M.*

tuberculosis would cause a shift in the gut microbiota. Here, we present a longitudinal survey using 16S ribosomal gene sequencing of the gut microbiota in a mouse model for TB. We assessed bacterial community composition and diversity prior to infection with *M. tuberculosis* CDC1551 and throughout infection until death. Further, we evaluated the gut microbiota from additional mice infected by a different *M. tuberculosis* strain (H37Rv) for comparison to the longitudinal study.

2.3 Materials and Methods

2.3.1 Ethics Statement

This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The Animal Protocol MO11M120 was approved by the Institutional Animal Care and Use Committee of the Johns Hopkins University, and covers all animal procedures, including time-to-death. All efforts were made to minimize suffering.

2.3.2 Bacterial Strains

Mycobacterium tuberculosis H37Rv and CDC1551 were routinely grown in Middlebrook 7H9 broth (Fisher Scientific, Waltham, MA) supplemented with 0.5% glycerol (Sigma, St. Louis, MO), 0.05% Tween 80 (Sigma) and 10% oleic acid-albumin-dextrose-catalase (OADC; Fisher Scientific) at 37^{o}C with agitation.

2.3.3 Animals

For the *M. tuberculosis* CDC1551 experiment, 4-6 week old female Balb/c mice were purchased from Charles River (Wilmington, MA). The five mice used for stool collection were toe tattooed for identification purposes and housed in the same cage. Mice were monitored daily as

part of a time-to-death experiment and were provided with gel cups broken up feed biscuits when they became cachectic. For the *M. tuberculosis* H37Rv experiment, 19-20 gram female Balb/c mice were purchased from Charles River. Mice were housed in cages containing 5 animals. For each group, 3 mice were randomly selected from one cage, and 2 from another, for stool collection. For colony forming unit (CFU) counts of lungs and spleens, mice were euthanized by cervical dislocation.

2.3.4 Stool Collection and Storage

For stool collection, each mouse was temporarily placed alone in a clean container until there was approximately 0.15 g stool in the container. The mouse was removed and the stool collected and stored in O-ring sealed tubes (Simport, Beloeil, Canada) at -80^{o}C.

2.3.5 Aerosol Infection

Mice were infected with log-phase broth cultures diluted in phosphate-buffered saline (PBS) using the Middlebrook inhalation exposure system (Glas-Col, Terre Haute, IN) in a single run. Bacterial burden was determined by sacrificing 3 mice per time point and enumerating CFUs on selective 7H11 plates (VWR, Radnor, PA).

2.3.6 DNA Extraction and 16S rRNA Sequencing

Total DNA was extracted from 0.15 grams of stool using the ZR Fecal DNA isolation kit (ZYMO Research Corp., Irvine, CA) with modifications including an enzymatic pre-treatment with mutanolysin (5000 units ml^{-1}, Sigma) and lysozyme (100 mg ml^{-1}, Sigma) in conjunction with aggressive bead beating using 0.1mm glass beads (BioSpec, Bartlesville, OK) and a bead-beater (BioSpec). Barcoded primers [27] were used to amplify the bacterial 16S rRNA gene region 27F-338R from 50 ng of purified DNA using AccuPrime High Fidelity DNA polymerase (Invitrogen,

Grand Island, NY) in a total reaction volume of 25 µL. Reactions were run in a PTC-100 thermal controller (MJ Research, Waltham, MA) using the following cycling parameters: 5 min of denaturation at 95 °C, followed by 25 cycles of 30 s at 95 °C, 30 s at 55 °C, and 60 s at 68 °C, with a final extension at 72 °C for 7 min. Amplicons were quantified using the Quant-iT PicoGreen dsDNA assay and equimolar amounts (100 ng) of the PCR product were mixed in a single tube. Amplification primers and reaction buffer were removed from each sample using the AMPure Kit (Agencourt, Brea, CA). The purified amplicon mixtures were sequenced by 454 FLX pyrosequencing using 454 Life Sciences primer A by the Genomics Resource Center at the Institute for Genome Sciences, University of Maryland School of Medicine.

2.3.7 16S rRNA sequence processing and analysis

Sequences generated from pyrosequencing of bacterial 16S rRNA gene amplicons from all mouse experimental groups were processed using mothur, version 1.27, according to the standard pipeline outlined in [28,29]. Sequences were denoised, trimmed, quality and chimera checked using *de novo* uchime, and clustered into operational taxonomic units (OTUs) at 97% pairwise identity and aligned to the Silva reference alignment [30]. Taxonomic classifications were assigned using the naïve Bayesian classifier with the May 2011 release of the greengenes database [31]. Rarefied OTUs (randomly subsampled to normalize sequence counts) were used to calculate community diversity for each sample. Unweighted and weighted UniFrac distances, a phylogenetically-sensitive measure of beta-diversity, was used as input for the Principle Coordinate Analysis (PCoA), with visualizations performed using the R package vegan [32]. An analysis of molecular variance (AMOVA) was performed to test whether the groups were statistically significant. In addition, the program Metastats was used to identify differentially abundant OTUs between pre- and post-infection groups [33]. Network analysis to evaluate how

the OTUs were partitioned among samples was performed using the make_otu_network.py script from QIIME [34] and visualized using Cytoscape [35].

2.3.8 Nucleotide accession numbers

The 16S rRNA 454 pyrosequencing data has been deposited in the GenBank Sequence Read Archive under accession number SRA060942.

2.4 Results

2.4.1 Compositional changes in the gut microbiota during *M. tuberculosis* infection

To study the impact of *M. tuberculosis* infection on the gut microbiota, we infected Balb/c mice with *M. tuberculosis* CDC1551, and monitored them until death. We collected fecal samples at time points prior to infection (pre-infection) and throughout infection (post-infection), selecting for analysis three pre-infection samples as controls, as well as samples from the first two weeks post-infection, the last two weeks prior to death and once per month in between (Figures 2-1, 2-2a, and Table 2-1). The fecal microbiota was characterized by 454 pyrosequencing of bacterial 16S rRNA gene amplicons (V1-V2 region) from the five infected mice. A total of 297,156 high-quality sequences were generated, corresponding to an average 6,322 reads per sample with an average length of 250 base pairs. These sequences were clustered into operational taxonomic units (OTUs) at 97% pairwise identity and taxonomically classified using the greengenes database [31]. The most abundant genera are shown in Figure 2-2b.

We further analyzed overall community diversity using the Shannon diversity index, a common ecological diversity measure, which takes into account both the number of species

(OTUs) present and their relative abundance (Figure 2-2c). There was an initial decrease in diversity in all mice post-infection, followed by a recovery in diversity until death or one week prior to death. This was true even for mouse 2, which survived 73 days longer than any of the other mice. These trends were also observed using the Inverted Simpson diversity index, which takes into account community richness, abundance, and is less sensitive to rare OTUs compared to the Shannon diversity index (Figure 2-3).

2.4.2 Gut community composition and structure differ based on infection status

To identify samples with similar microbial community structure and composition, we implemented multidimensional cluster analysis based on the weighted and unweighted UniFrac distances (Figures 2-4a and 2-4b). UniFrac is a phylogenetically-aware measure of beta-diversity that can be used to compare OTU structure and community diversity. The weighted measure takes into account phylogenetic tree branch length. Both measures showed clear clustering among the uninfected samples taken pre-infection and the infected samples collected post-infection. We utilized an analysis of molecular variance (AMOVA), a statistical model similar to analysis of variance that is used to analyze differences in genetic diversity, to test whether the pre-infection and post-infection samples were statistically different, and found both weighted and unweighted UniFrac measures were significantly different ($p < 0.001$). We further utilized a network analysis to evaluate how the OTUs were partitioned among samples, lending additional support for the separation of pre- vs. post-infection samples. Strikingly, the first two weeks after infection were found to be intermediate between the samples taken before infection and later in infection, with the Firmicutes distinguishing between the two conditions (Figure 2-4c)

We next implemented a categorical analysis to evaluate whether specific bacterial OTUs discriminated between pre-infection and the infected samples collected post-infection. Eighty-eight OTUs were found to be significantly differential ($q<0.01$); the majority belonged to the Firmicutes, specifically within the order Clostridiales (Figure 2-5, Supplemental Table 2-S1) [33]. The bulk of these OTUs classified within the Lachnospiraceae family, 31 affiliated with the unclassified Lachnospiraceae OTU2087 clade, and the Ruminococcaceae family. Interestingly, all of these significantly different OTUs were more abundant pre-infection compared to post-infection.

2.4.3 Distinct changes in the gut community is independent of *M. tuberculosis* strain

To determine if our previous results could be replicated with a different *M. tuberculosis* strain, we collected fecal samples from a single time point (46 days post infection) from *M. tuberculosis* H37Rv infected Balb/c mice and age-matched uninfected Balb/c mice that had been in the same facility for the same amount of time. Each group of mice was kept in two different cages to rule out variations in caging conditions. As described for the previous study, the fecal microbiota was characterized by 454 pyrosequencing of bacterial 16S rRNA gene amplicons (V1-V2 region) for this set of mice. A total of 148,466 high-quality sequences were generated, corresponding to an average 14,847 reads per sample, and were analyzed with the first experiment dataset to maintain consistency in OTU clustering (see Methods). The most abundant genera are shown in Figure 2-6. We implemented multidimensional cluster analysis based on the weighted and unweighted UniFrac distances and found distinct clustering between infected and uninfected samples, congruent with the results for the longitudinal study (Figures 2-7a and 2-7b). Network analysis similarly confirmed these observations (Figure 2-7c). These

results corroborate our earlier observations of distinct clustering between infected and uninfected animals.

Additionally, we found considerable overlap in the phylogenetic affiliation of the differentially abundant OTUs, which discriminated between uninfected and H37Rv-infected samples, compared to the CDC1551 infection longitudinal study. Seventy-three OTUs were found to be differential using the categorical analysis ($q<0.01$), the vast majority similarly within the Lachnospiraceae and Ruminococcaceae families (Figure 2-8 and Supplemental Table 2-S2). Remarkably, the phylogenetic affiliations of the discriminatory OTUs mirrored those found in the CDC1551 infection longitudinal study and were highly abundant in the uninfected group, including OTU2087 (family Lachnospiraceae), OTU1995 (order Clostridiales), OTU1998 (genus *Catabacteriaceae*), OTU2159 (family Ruminococcaceae), OTU2166 (genus *Clostridium*), and OTU2176 *(*genus *Oscillospira*). However, while there was high congruence in the phylogenetic lineages observed between the two *M. tuberculosis* studies, only five OTUs specifically overlapped in both experiments (Appendices A and B).

2.5 Discussion

We have monitored changes in the composition of the mouse gut microbiota from pre-infection to death, and shown that there is a clear difference between the microbial communities of infected and uninfected mice, results that have been confirmed by the use of two different strains of *M. tuberculosis.* These changes occur by day six post aerosol infection, indicating that this shift is very rapid. The prompt shift in the community after aerosol infection suggests that the gut microbiota is modulated during TB infection and may be responding to host immunological changes [36]. Furthermore, there is a consistent trend in these alterations, with a decrease in overall community diversity and discrete shifts in specific microbial taxa,

despite the difference in survival time for the five mice studied. This trend occurred even in mouse 2, which survived 73 days longer than the other mice. Overall, we identified a number of OTUs significantly different between infected and uninfected mice, regardless of the infecting strain of *M. tuberculosis*. Mycobacterial DNA was not detected in any of these samples with the 16S rRNA pyrosequencing, but PCR of the *IS6110* insertion sequence, which is specific for *M. tuberculosis*, in the last sample collected prior to death identified TB in only three of the five mice (data not shown). We posit that the observed microbiota changes are not be mediated directly by the presence of *M. tuberculosis* in the gut, but instead represent crosstalk between the resident microbiota and the mucosal immune system. In fact, the minimum in diversity approximately corresponds to the time when the adaptive immune system begins to achieve a plateau in bacterial burden [37]. Thus, the loss of diversity could be a result of immune system activation, with a recovery once bacterial burden and immune activity have reached equilibrium.

We observed significant differences in the relative abundance of members within the Lachnospiraceae and Ruminococcaceae families (Clostridiales) in the two studies, with higher relative abundances in both the uninfected and pre-infected mice. It is known that some members of clusters IV and XIVa of the genus *Clostridium* (a genus in the *Clostridiales*) induce regulatory T cells [38]. Members of the Lachnospiracaceae have also been shown to decrease in abundance in inflammatory bowel disease (IBD) and colitis [39,40]. Furthermore, a recent study found that a decrease in abundance of OTU 2087 was correlated with exacerbated asthma [41]. Future work to delineate immune-modulatory members of the Clostridiales, particularly those members associated with regulatory T cells and which have a role in *M. tuberculosis* immunity, will be critical.

While not as prominent as the differential *Clostridiales* OTUs, we additionally observed a subset of *Bacteroidales* OTUs differentially abundant in pre-infected and uninfected samples in the two studies. *Bacteroides* spp. have been shown to play important anti-inflammatory roles in inhibiting activation of the NF-κB pathway and induction of IL-10-producing T cells [15]. Furthermore, *Bacteroides fragilis* has been shown to modulate the T-helper type 1/2 (Th1/Th2) balance, another aspect of the immune system critical for the control of *M. tuberculosis* [42]. Further studies to characterize the relationship of specific *Bacteroides* species to the immune system are clearly needed.

Together, our results establish that the gut microbiota of mice changes significantly following aerosol infection by *M. tuberculosis,* and these differences may be related to the immune signaling from lung to gut. These changes begin as early as six days post infection and are characterized by an initial loss in diversity, which recovers to a significantly different composition. During this time, there are many alterations in the relative abundance of a number of OTUs, most significantly a decrease in members of the *Clostridiales* and *Bacteroidales*. These observations have important implications for our understanding of the interplay between the immune system and the gut microbiota. Further investigation into the interaction of the host immune system and the gut microbiota could lead to a more mechanistic understanding of gut immune functioning and the role the microbial fraction plays during TB infection.

2.6 Figures

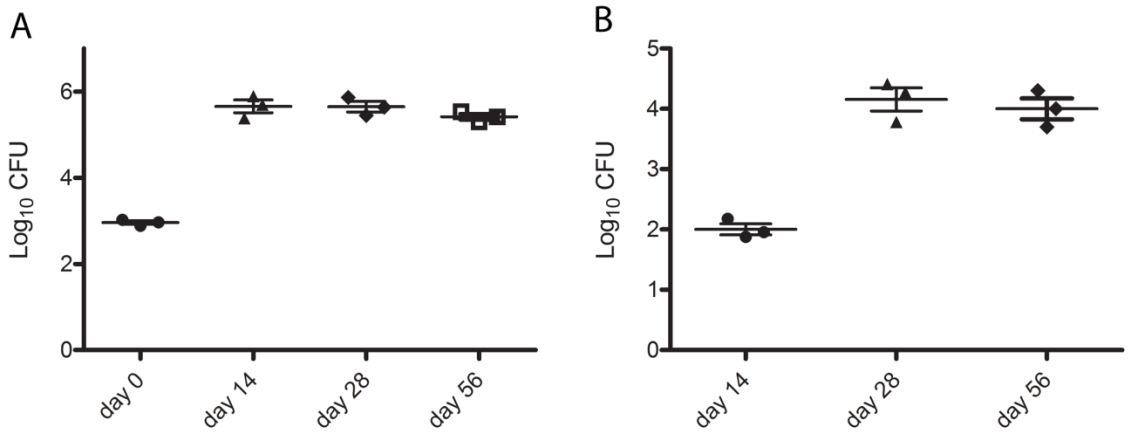


Figure 2-1. Bacterial burden of *M. tuberculosis* CDC1551 infected mice.

M. tuberculosis colony forming units (CFUs) at day 0, 14, 28 and 56 in (**A**) the lungs and (**B**) the spleen of mice infected at the same time as the mice followed to death.

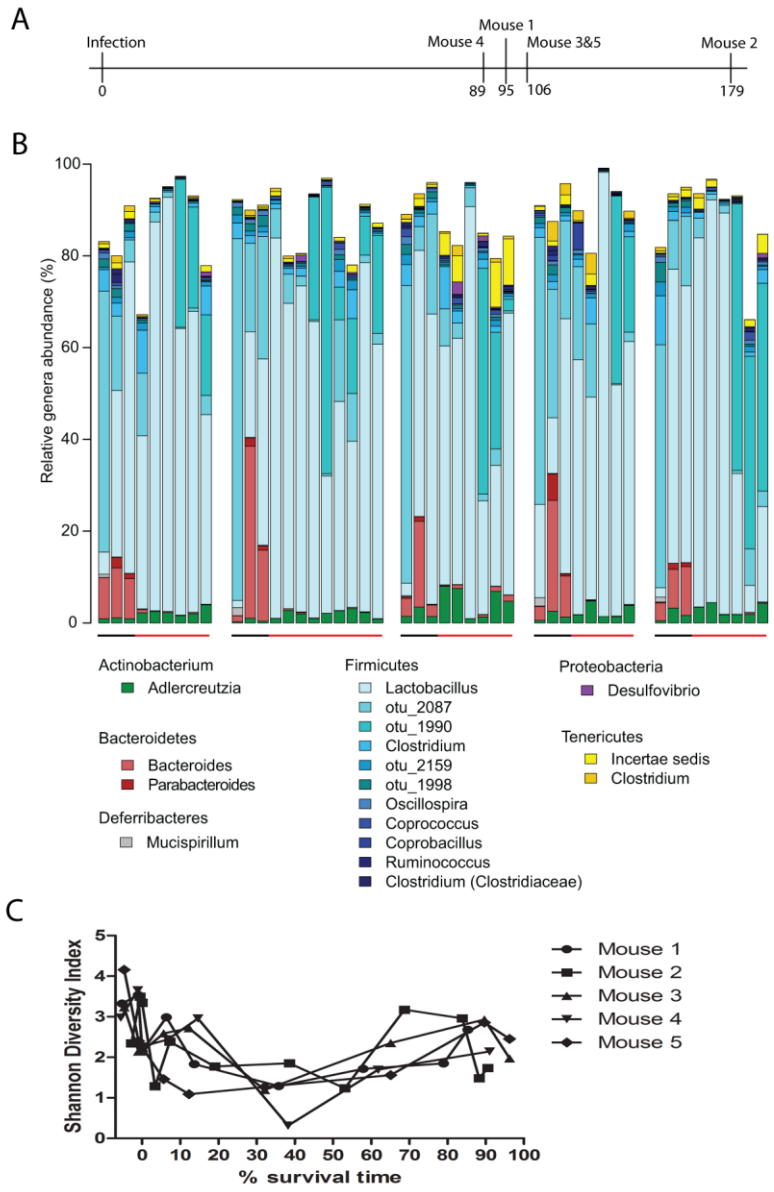


Figure 2-2. Community structure of individual *M. tuberculosis* CDC1551 infected mice over time.

(**A**) Survival time in days post-infection for each mouse. (**B**) Phylogenetic profile of bacterial genera. Stacked bar charts in chronological order for each mouse of the 18 main genera identified based on ≥ 1% abundance present in at least two samples. Unclassified sequences are not shown. Black colored bars along x-axis indicate samples taken prior to infection, while red

colored bars indicate post-infection. Each group represents an individual mouse, followed to death. The mice are represented sequentially, with mouse 1 on the left, and mouse 5 on the right. (**C**) Community diversity in each sample as measured by the Shannon diversity index, plotted against the percent survival time.

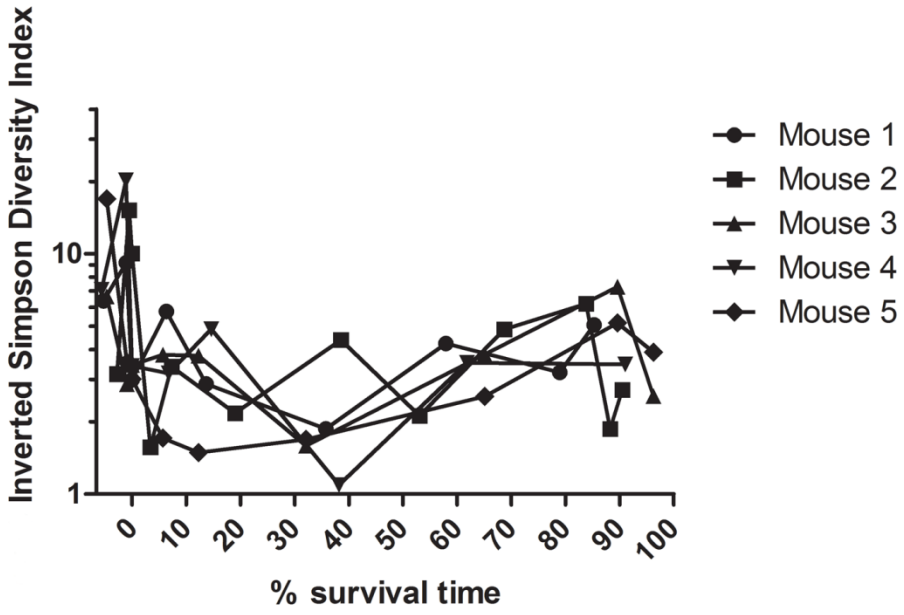


Figure 2-3. Community diversity of *M. tuberculosis* CDC1551 infected mice.

Community diversity in each sample as measured by the Inverted Simpson diversity index, plotted against the percent survival time.

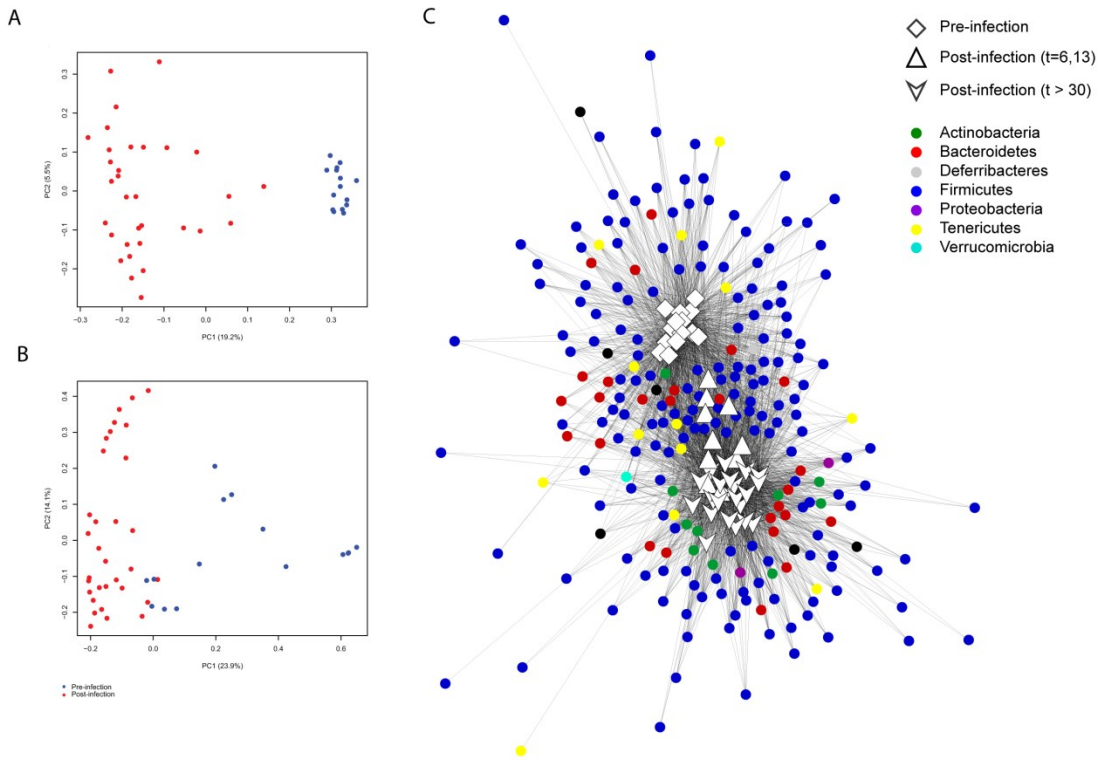


Figure 2-4. Composition of the gut microbiota significantly changes with *M. tuberculosis* CDC1551 infection.

(**A**) Unweighted and (**B**) weighted Unifrac measures of beta-diversity visualized using Principle Coordinate Analysis (PCoA) following individual mice over time with *M. tuberculosis* CDC1551 infection. Blue dots indicate samples collected pre-infection. Red dots indicate samples collected post-infection. Variance for first two component axes is shown as percent of total variance. An analysis of molecular variance (AMOVA) was performed to test whether the separation of uninfected and TB-infected samples was statistically significant. In both unweighted and weighted Unifrac measures, there was a statistically significant difference ($p<0.001$). (**C**) Network analysis of OTUs partitioned among samples, using a five sequence cutoff, and colored by phylum.

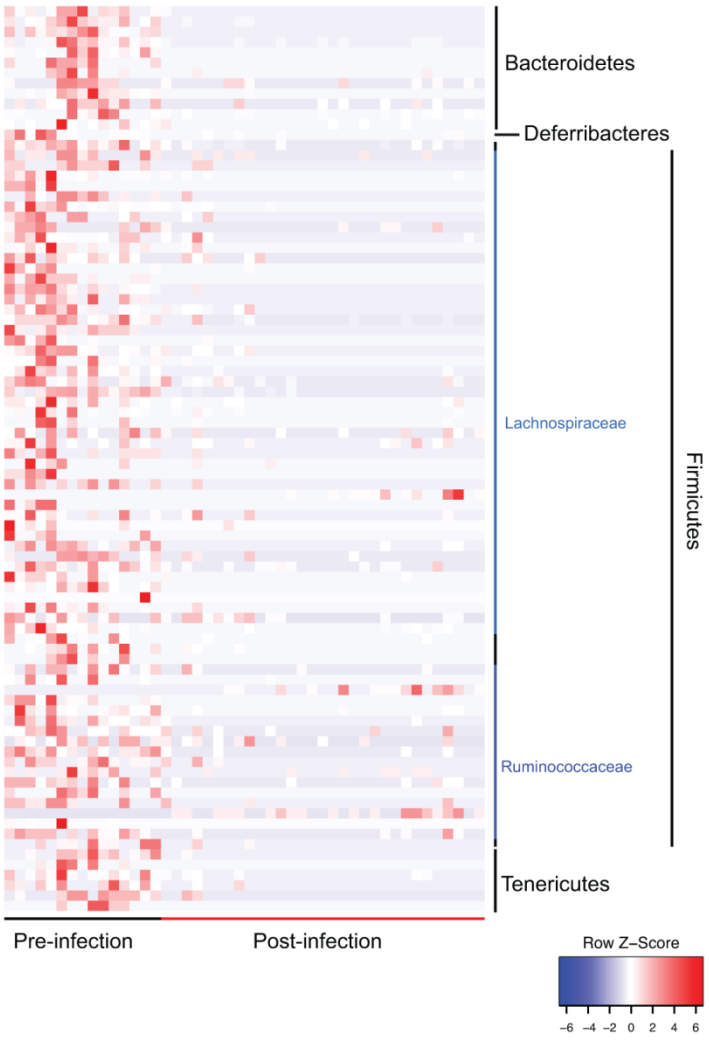


Figure 2-5. Differentially abundant OTUs identified between pre-infection and post-infection.

OTUs are ordered by consensus taxonomic classification, with OTUs scaled by relative abundances for each row ranging from low relative abundance (blue) to high relative abundance (red).

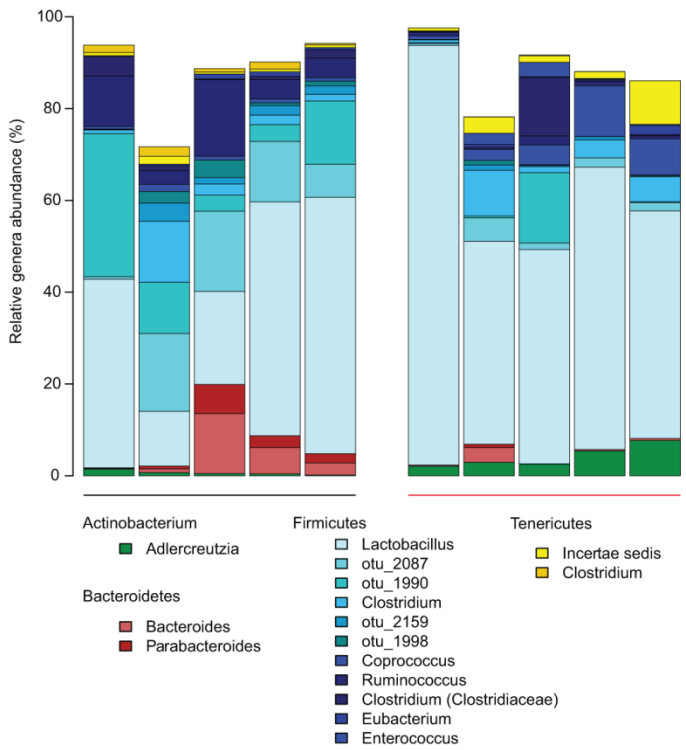


Figure 2-6. Phylogenetic profile of bacterial genera for uninfected and *M. tuberculosis* H37Rv infected mice.

Stacked bar charts for uninfected and H37Rv-infected mice of the 16 main genera identified based on ≥ 1% abundance present in at least two samples. Unclassified sequences are not shown. The black colored bar along x-axis indicates the five uninfected mice, while the red colored bar indicates mice infected with H37Rv.

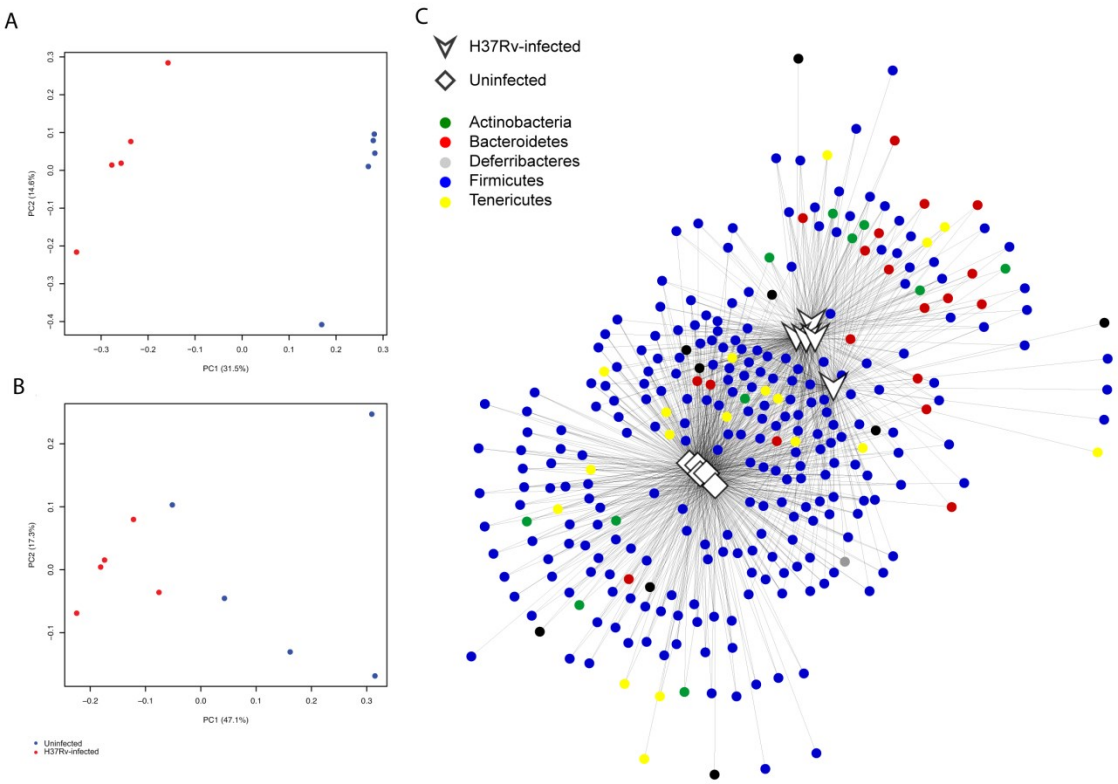


Figure 2-7. Gut microbiota composition of *M. tuberculosis* H37Rv infected mice is significantly different from uninfected mice.

(**A**) Unweighted and (**B**) weighted Unifrac measures of beta-diversity visualized using Principle Coordinate Analysis (PCoA) for the comparison of H37Rv-infected mice to uninfected mice at a single time point. Blue dots indicate samples collected pre-infection. Red dots indicate samples collected post-infection. Variance for first two component axes is shown as percent of total variance. In both unweighted and weighted Unifrac measures, there was a statistically significant difference (AMOVA $p \leq 0.005$). (**C**) Network analysis of OTUs partitioned among samples, using a five sequence cutoff, and colored by phylum.

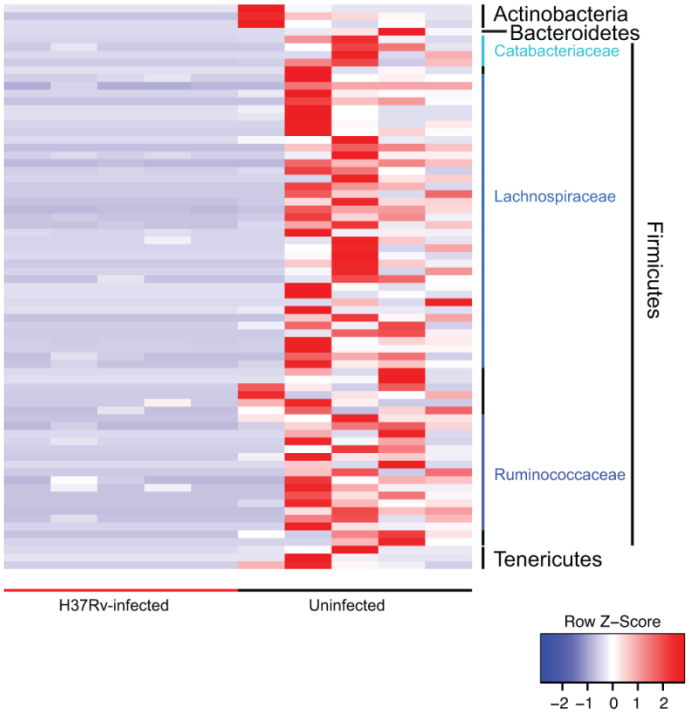


Figure 2-8. Differentially abundant OTUs identified between uninfected and *M. tuberculosis* H37Rv infected mice.

OTUs are ordered by consensus taxonomic classification, with OTUs scaled by relative abundances for each row ranging from low relative abundance (blue) to high relative abundance (red).

2.7 Table

Mouse ID	Time points sampled (days)	Time of Death (day)
1	-5, -1, 0, 6, 13, 34, 55, 75, 81	95
2	-5, -1, 0, 6, 13, 34, 69, 95, 123, 150, 158, 162	179
3	-5, -1, 0, 6, 13, 34, 69, 95, 102	106
4	-5, -1, 0, 6, 13, 34, 55, 81	89
5	-5, -1, 0, 6, 13, 34, 69, 95, 102	106

Table 2-1. Time points assayed and time of death for each mouse.

Stool samples were collected every week post-infection. From these specimens, we selected samples from the first two weeks post-infection, the last two weeks prior to death and once per month in between for sequencing analysis. Time is given in days relative to infection.

Chapter 3: Interaction between *M. tuberculosis*, the immune system, and the microbiota results in specific changes and is mediated by the adaptive immune system

3.1 Abstract

Previous work has shown a rapid and significant change in the gut microbiota of mice aerosol infected by *Mycobacterium tuberculosis*. We were interested in the mechanism behind these changes and hypothesized that the immune system plays a critical role in this response. To determine the mechanism, we infected Balb/c, Black/6, MyD88$^{-/-}$ and RAG$^{-/-}$ mice with *M. tuberculosis* and assessed changes in species and gene content in the gut microbiota through 16S rDNA sequencing, whole genome sequencing and RNA-seq. We identified organisms and KEGG orthologs significantly associated with *M. tuberculosis* infection, as well as with changes in cytokine levels. Furthermore, we confirmed our previous findings, showing that the microbiota changes by day 10 post-infection in multiple mouse genotypes. In addition, we found that the microbiota of RAG$^{-/-}$ responds differently than the other genotypes, indicating that the changes detected are mediated through the adaptive immune system. Finally, we demonstrated a significant difference between the microbiota of *M. tuberculosis* infected samples and *M. avium* or *M. smegmatis* infected samples, suggesting that these changes are specific to *M. tuberculosis*. Taken together, we have shown that the gut microbiota of mice responds specifically to aerosol *M. tuberculosis* infection through changes mediated by the adaptive immune system.

3.2 Introduction

The role of the microbiome in *Mycobacterium tuberculosis* infection remains largely unexplored. As discussed in chapter 2, prior to the start of our work, only three papers had been published on *M. tuberculosis* and the microbiota. We showed that the gut microbiota undergoes significant changes in response to aerosol *M. tuberculosis* infection. Since that time, in addition to our work, several additional papers have been published. One paper showed that the microbiota plays an important role in the efficacy of vaccines, including BCG, the vaccine for tuberculosis [43]. Another paper looked at the microbiota of the sputum, oropharynx and nasal respiratory tract of pulmonary tuberculosis patients and healthy controls and found changes with infection, particularly in the oropharynx [44]. They also found some significant changes in the fungi of the microbiota. A third paper compared the sputum of new, recurrent and treatment failure tuberculosis cases to throat swabs from healthy controls [45]. Like us, they found loss of diversity with tuberculosis infection. These last two papers both identified changes at the phyla level that were associated with *M. tuberculosis*. Our results were more specific, but overlapped many of these phyla. Thus, there is a small but growing body of literature that supports our findings of changes in the microbiota associated with *M. tuberculosis* infection.

Given the rapid changes we observed in our previous study, we hypothesized that the changes we saw were mediated by the immune system. Many studies have shown an interaction between the immune system and microbiota (reviewed in [46,47]). These studies have demonstrated that the presence of particular aspects of the microbiota influence development of both the innate and adaptive immune system, and dysbiosis (alterations in the community) results in disease [38,48,49].

As a result, we were interested in the role the immune system plays in mediating the interaction between the microbiota and *M. tuberculosis*. To address this issue, we studied mouse strains lacking either the innate or the adaptive immune system and compared them to wild-type strains. We analyzed changes in the microbial composition in terms of both bacterial species and gene content through 16S rDNA sequencing, whole genome sequencing and RNA-seq. We confirmed our previous findings of changes in the microbiota in response to *M. tuberculosis* infection and showed that these changes are specific to *M. tuberculosis*. Furthermore, we demonstrated that the microbiome of mice lacking B and T cells responds differently to the microbiome of other mice, suggesting that the adaptive immune system mediates these changes.

3.3 Materials and Methods

3.3.1 Bacterial Strains

Mycobacterium tuberculosis CDC1551, *Mycobacterium smegmatis* mc^2155 and *Mycobacterium avium* were routinely grown in Middlebrook 7H9 broth (Fisher Scientific, Waltham, MA) supplemented with 0.5% glycerol (Sigma, St. Louis, MO), 0.05% Tween 80 (Sigma) and 10% oleic acid-albumin-dextrose-catalase (OADC; Fisher Scientific) at 37^{o}C with agitation.

3.3.2 Animals

4-6 week old female Balbc/J, C57BL/6J, *Rag1*^{tm1Mom}, and B6.129P2(SJL)-*Myd88*$^{tm1.1Defr}$/J mice were purchased from Jackson Laboratory (Bar Harbor, ME). Mice used for stool collection were toe tattooed for identification purposed and each group of 5 mice was housed in two separate cages.

3.3.3 Stool collection and storage

For stool collection, each mouse was temporarily placed alone in a clean container until there was approximately 0.15 g stool in the container. The mouse was removed and the stool collected and stored in O-ring sealed tubes (Simport, Beloeil, Canada) with 1mL RNA*later* (Life Technologies, Grand Island, NY) at 4 ^{o}C overnight. RNA*later* was then removed and stool was stored at -80^{o}C.

2.3.4 Aerosol Infection

Mice were infected with log-phase broth cultures of bacteria diluted to OD 0.15 in phosphate-buffered saline (PBS) using the Middlebrook inhalation exposure system (Glas-Col, Terre Haute, IN). Bacterial burden was determined by sacrificing mice per Table 3-1 and enumerating CFUs. *M. tuberculosis* and *M. avium* were grown on selective 7H11 plates (VWR, Radnor, PA), while *M. smegmatis* was grown on 7H10 plates (VWR, Radnor, PA).

2.3.5 Determination of cytokine levels

Lungs or spleens were harvested from mice and put into O-ring tubes containing 1mL PBS and immediately flash-frozen in liquid nitrogen then stored at -80^{o}C. Organs were thawed on ice and bead-beaten three times for 30 seconds with 2mm glass beads. Supernatent was filtered through 0.22μm cellulose acetate spin filters (Costar). Samples were run on a Luminex (luminex Corporation, Austin, TX) using a custom Milliplex kit from EMD Millipore Corporation (Billerica, MA).

2.3.6 RNA and DNA isolation

Two stool pellets were washed with 1mL ice cold PBS(Cellgro), then 300μL of TE-SDS (Sigma, St. Louis, MO), 500μL of TE-saturated phenol (Sigma), and 0.3g of 0.1mm glass beads

(BioSpec, Bartlesville, OK) was added. Samples were bead-beaten (Biospec) for 2 minutes then spun at 12,000 rpm for 5 minutes at 4°C. Phenol-chloroform-isoamyl alcohol (Sigma) was added to the aqueous phase in a new RNase-free tube (Ambion) and the sample was spun at 12,000 rpm for 5 minutes at 4°C. The aqueous phase was moved to a new RNase-free tube and nucleic acid was precipitated with 3M sodium acetate (Ambion) and isopropanol (Sigma). The sample was spun at 10,000 rpm for 5 minutes at 4°C and the pellet washed with 70% ethanol, spun, and resuspended in RNase-free water. The concentration was measured on a Nanodrop and diluted to at most 50µg in 50µL. RNA and DNA were then separated and purified using the QIAgen Allprep RNA/DNA mini kit (Qiagen).

2.3.7 Sequencing

16S rDNA sequencing and whole genome sequencing were performed as described in [50] while RNA-seq was performed as described in [51].

2.3.8 Nucleotide accession numbers

Data has been deposited in GenBank as a BioProject and can be accessed using accession number PRJNA219721 (http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA219721).

2.3.9 16S rDNA analysis

Reads were processed as described in [50] and OTU picking was performed using QIIME 1.6.0 with the may2013 greengenes database. Gene content predictions were made using PICRUSt [52]. For comparisons with the data from [53], all reads were run through the RDP classifier version 2.8 [31].

2.3.10 Whole genome sequencing analysis

Reads were cleaned as described in [50] and then run through HUMAnN [54]or MetaPhlAn v2.0 [55].

2.3.11 RNA-seq analysis

Reads were cleaned as described in [51] and run through HUMAnN [54]. Genes were sum-normalized and then Witten-Bell smoothing was performed [51]. Genes that did not have an abundance of at least 0.00001 in at least two samples were filtered, and then a log transformed ratio of the RNA counts to the counts from the HUMAnN whole genome sequencing reads was calculated. These ratios were used for all subsequent analyses. Limma was used to identify changes in expression compared to day -3.

2.3.12 Identification of associations

Associations between infecting organism and relative abundance were determined using MaAsLin [56], controlling for cage. NMDS plots were generated using BreadCrumbs (https://bitbucket.org/biobakery/breadcrumbs). P-values on the NMDS plots were calculated on sum-normalized data using the adonis method in vegan [32] using Euclidean distances, and adjusted using the p.adjust in R.

3.4 Results

3.4.1 Some OTUs are correlated with *M. tuberculosis* burden and cytokine level, regardless of host genetics

Given the previous observations of rapid changes induced in the gut microbiota after aerosol infection with *M. tuberculosis*, we hypothesized that the immune system is mediating

the changes in the microbiota [53]. In order to address this hypothesis and analyze the role of the immune system in this interaction, we focused on four mouse genotypes (Table 3-1). One was *Rag1*^{tm1Mom} (hereafter abbreviated RAG$^{-/-}$). Rag1 (recombination-activating gene) is a crucial part of VDJ recombination and the development of B and T cell receptors, and as a result, these mice lack both B and T cells, and thus a functional adaptive immune system [57]. The second mouse strain we studied was B6.129P2(SJL)-*Myd88*$^{tm1.1Defr}$/J (hereafter abbreviated MyD88$^{-/-}$). MyD88 is an adaptor protein critical for signaling through Toll-like receptors, and as a result these knockout mice lack an innate immune system. [57]. Both of these mice were developed on a C57BL/6J background, and so the wild-type strain (hereafter referred to as Black/6) was included as a control. Finally, we included Balbc/J (hereafter referred to as Balb/c) as a positive control because this strain was used in our previous studies (chapter 2, [53]). Mice were either infected with *M. tuberculosis* CDC1551, or left as uninfected controls. In addition, to address whether the changes we saw were specific to *M. tuberculosis*, we infected three Balb/c mice with *M. avium* and three Black/6 mice with *M. smegmatis*. *M. avium* is a slow growing opportunistic mycobacterial pathogen while *M. smegmatis* is a non-pathogenic fast-growing mycobacterial strain frequently used to study mycobacterial genomics [58,59]. Figure 3-1 shows how these different bacteria grow in the different mouse genotypes.

Stool samples were collected from pre-infection to one month post-infection (Table 3-1). By three weeks, the adaptive immune had started to respond to *M. tuberculosis* infection in wild-type mice (as evidence by the plateau in bacterial growth in Figure 3-1) and by four weeks all infected knock-out mice were close to death and had to be sacrificed early. We isolated RNA and DNA from these stool samples and performed 16S rDNA sequencing, whole genome sequencing, and RNA-seq (Table 3-1). From this data, we were able to determine which

organisms were present in the microbiota as represented by OTUs (operational taxonomic units), the functional content of the microbiota, and the changes in gene expression.

We first wanted to know whether there were any correlations between the relative abundance of a particular OTU or functional category and the amount of *M. tuberculosis* present in the mouse (represented by colony forming units, or CFUs). Thus, we took the average of the OTUs or KEGG orthologs present at each timepoint with CFU counts and used MaAsLin to identify any correlations. MaAsLin is a tool developed by the Huttenhower lab that can identify associations between metadata and microbial species or gene abundances [56]. Supplemental Table 3-S1 lists the significant OTUs and KEGG orthologs associated with *M. tuberculosis* CFU count. For both 16S and whole genome sequencing data, all OTUs significantly associated with bacterial burden in the spleen were also significantly associated with bacterial burden in the lungs, or had an adjusted p-value of 0.052 in the lung. Most of the significant OTUs from 16S rDNA sequencing were in the order Clostridiales, with the exception being the most significant OTU in the spleen, which was in order RF39 of the Mollicutes class. An OTU in the *Adlercreutzia* genera was the most significantly associated OTU in the lung (and was identified in the studies in chapter 2), but 21 Firmicutes and one Tenericutes were also significant in the lung. Likewise, for the OTUs identified through whole genome sequencing, all OTUs significantly associated with spleen CFU were also significantly associated with lung CFU. Unlike for 16S, these OTUs were all in the phyla Actinobacteria, particularly *Enterorhabdus caecimuris*. Spleen forming virus was also associated with lung bacterial burden. These OTUs may differ from the 16S rDNA sequencing results because we identified fewer OTUs in our whole genome sequencing data (see below). Finally, three KEGG orthologs were associated with CFU burden in the lungs when analyzing DNA – a transporter, an aminopeptidase and a permease (Supplemental Table 3-S1). When taking gene expression into account through RNA-seq, two different KEGG orthologs were associated

with CFU burden in the lungs – a hydrolase and an unannotated group (Supplemental Table 3-S1). Taken together, this suggests that specific elements of the microbiota are changing in response to *M. tuberculosis* infection, regardless of mouse genotype.

In addition to the bacterial burdens of these mice, we also determined the amount of IFNγ, IL-10, IL-12(p40), TNFα, IL-17, IL-1β, IL-2 and IL-6 in the lungs and spleen (Figure 3-2). Similar to our analysis of the correlation between *M. tuberculosis* CFU and the microbiota, we used MaAsLin to identify OTUs or KEGG orthologs associated with cytokine levels (Supplemental Table 3-S2). Very few OTUs were significantly correlated with any cytokine: an OTU in the order Clostridiales and two viruses were associated with IL-10 levels in the lungs, and an *Enterococus* genera OTU was associated with IFNγ levels in the lungs. In contrast, in the DNA analysis two KEGG orthologs associated with lung IFNγ levels, eleven with lung IL-10, and 28 with spleen IL-1β. In the RNA-seq analysis, three KEGG orthologs were associated with lung TNFα levels, one with spleen IL12p40 and one with spleen IFNγ. Thus, we find evidence across mouse genotypes of the microbiota responding to certain aspects of the immune system.

3.4.2 Correlation between OTU composition from 16S rDNA sequencing, mouse genotype, and mycobacterial infection

We were first interested in how this experiment compared to our previous analyses (chapter 2, [53]). Several parameters were different between this experiment and previous work. First, in order to obtain the knock-out mice, we changed vendors from Charles River Laboratories to the Jackson Laboratory. The source of mice has previously been shown to have an effect on microbial composition [60]. In addition, in order to isolate RNA for RNA-seq, we changed our nucleic acid isolation protocol (see methods). Finally, a different 16S variable region was sequenced. To determine how these changes affected the microbiota, we compared

our previous data and the 16S data from the Balb/c mice from this experiment. We used RDP to pick OTUs for all samples and then graphed them on a NMDS plot (Figure 3-3)[31]. The samples from chapter 2 (Experiment 1 and 2) separated from the samples from this experiment (Experiment 3). This indicated that our Balb/c baseline was significantly different, making it difficult to compare directly to our previous results.

This difference was also evident when we look at the OTU composition of our samples (Figure 3-4). Compared to our previous studies, the differences between *M. tuberculosis* infected and uninfected Balb/c samples was no longer as clear. However, our new data did confirm that there are differences in the microbial composition of the different mouse phenotypes, such as the high proportion of Verrucomicrobia in uninfected $RAG^{-/-}$ samples compared to uninfected samples from the other backgrounds. Interestingly, this OTU increases in abundance with infection of Black/6 and $MyD88^{-/-}$ mice and is significantly associated with *M. tuberculosis* infection in Black/6 mice (Supplemental Table 3-S3).

Caging effects have been reported in mouse microbiota studies, although we did not see them in our previous work [61,62]. However, since this could provide a major confounding effect in this experiment, we separated each group of five mice into two cages. We looked at the pre-infection timepoints for all mice, separating them by mouse genotype, and found no significant caging effect, with the nonsignificant exception of the Black/6 mice later infected by *M. smegmatis* (Figure 3-5). This was also true for the whole genome sequencing and functional content analyses (data not shown). Thus, in these experiments, we did not have a caging effect.

When we compared all samples together, we found that most variation was in mouse genotype (Figure 3-6), with the Balb/c samples separating from the other genotypes (which were on a Black/6 background). As a result, we then looked at each genotype separately. In this

experiment, the difference between all *M. tuberculosis* infected and uninfected Balb/c samples was not significant (Figure 3-7a and b). However, from these plots, it appeared that much of the overlap was in the early post-infection timepoints (days 1-7). Thus, we separated the timepoints into early post-infection (first week), late post-infection(last week of experiment, when adaptive immune system has begun to respond and knockout mice are close to death) and middle (everything in between) (see Table 3-1). When we removed the early post-infection timepoints, the difference between infected and uninfected *M. tuberculosis* samples became significant, similar to our previous observations (Figure 3-7c and d). The delayed change compared to previous observations is most likely due to the altered microbial composition of these mice (Figure 3-4). Nevertheless, by day 10 the gut microbiota is significantly different between *M. tuberculosis* infected and uninfected Balb/c mice.

We next focused on the other mouse genotypes. Infected and uninfected Black/6 samples were significantly different, even with the early timepoints included (Figure 3-8). Likewise, MyD88$^{-/-}$ samples were significant with and without the early timepoints, suggesting that the innate immune system is not required to mediate this response (Figure 3-9). In contrast, RAG$^{-/-}$ samples were not significantly different with infection, suggesting that the changes do require the adaptive immune system (Figure 3-10).

Given these significant changes, for each genotype we looked at which OTUs were significantly different between uninfected and *M. tuberculosis* infected samples using MaAsLin (Supplemental Table 3-S3). No OTUs were significant in all genotypes (maroon section of Figure 3-11). However, 3 OTUs were significant in Balb/c, Black/6 and MyD88$^{-/-}$ with early timepoints removed, when all three mouse genotypes were significantly different (light blue section of Figure 3-11b). These three OTUs were an OTU in the *Adlercreutzia* genera in the Actinobacteria

phyla, an OTU in the order RF39 in the Tenericutes phyla, and an OTU in the family S24-7 in the Bacteroidetes phyla. The first was also significant when looking for correlations with *M. tuberculosis* CFU, and the second was very closely related to another OTU from that analysis. Thus, there appears to be a strong association between *Adlercreutzia* and *M. tuberculosis* infection.

3.4.3 Correlation between whole genome sequencing OTU composition, mouse genotype, and mycobacterial infection

In addition to OTU data from 16S rDNA sequencing, we obtained whole genome sequencing data for a subset of timepoints. We used MetaPhlAn to predict which OTUs were present [55]. Unfortunately, MetaPhlAn was developed for the human microbiome, and was only able to identify 150 unique OTUs in our dataset (Figure 3-12). This is in contrast to the 1,210 identified through 16S rDNA sequencing. As a result, the composition of these samples was very different from that predicted by 16S rDNA sequencing. For example, although MetaPhlAn gave us relative abundance data for viruses and eukaryotic organisms, which were missing from our 16S sequencing data, some bacterial phyla detected by 16S sequencing were missing from this analysis, such as the Tenericutes, some of which were associated with infection (Figure 3-4 and Figure 3-12). With these reduced samples and reduced number of OTUs, although there was still a significant difference between mouse genotypes, the difference between *M. tuberculosis* infected and uninfected samples was no longer significant for any mouse genotype, even with the early timepoints removed (Figure 3-13 to Figure 3-17). Furthermore, of the few OTUs significantly associated with infection status, none were significant in more than one genotype (Supplemental Table 3-S4, Figure 3-18). In addition, no OTUs were significant in both 16S rDNA and MetaPhlAn analysis (Supplemental Table 3-S3 and

3-S4). Thus, in our data set, 16S sequencing provides better resolution, although the whole genome sequencing data does follow a similar trend to the 16S. Removing early timepoints from Balb/c mice makes the difference between infected and uninfected more significant and Black/6 and MyD88$^{-/-}$ are almost significant while RAG$^{-/-}$ is not (Figure 3-14 to Figure 3-17). This corroborates our observations that the changes in the gut microbiota become more significant with time and that lack of an adaptive immune system eliminates these changes.

3.4.4 Correlation between microbiota functional content measured by 16S rDNA sequencing, mouse genotype, and mycobacterial infection

Given the changes in microbial composition with *M. tuberculosis* infection, we were interested in whether the functional capabilities of the microbiome also changes. We used PICRUSt to predict functional content from our 16S data. PICRUSt is a tool that uses OTUs identified through 16S rDNA sequencing to estimate the relative abundance of gene families [52]. When initially plotting PICRUSt results, 31 samples were clearly different from the others and appeared as outliers with much higher NMDS values than any other sample. 3 of these were from MyD88$^{-/-}$, 2 from RAG$^{-/-}$, and the rest from Balb/c. Since this was not seen in the 16S data, we concluded that they were artifacts of the prediction algorithm, perhaps because it was optimized for human and not mouse microbiota. Thus, these samples were removed from further analysis of PICRUSt data.

These data provided a different picture to the changes seen at OTU level (Figure 3-19 to Figure 3-24). Although the difference between genotypes is significant, this is largely due to differences between RAG$^{-/-}$ and the other genotypes (Figure 3-19). Balb/c was not significantly different between *M. tuberculosis* infected and uninfected samples, even with the early timepoints removed, but all other genotypes, including RAG$^{-/-}$, were significantly different,

especially with the early timepoints removed (Figure 3-20 to Figure 3-23). However, when looking at KEGG orthologs significantly associated with infection, $RAG^{-/-}$ had no significant features and Balb/c only had one, while Black/6 and $MyD88^{-/-}$ had 239 and 178 respectively, 175 of which were shared (Supplemental Table 3-S5, Figure 3-24). Thus, *M. tuberculosis* infection has an effect on gene content as predicted from 16S sequencing, but this change is different from that seen at the OTU level. Nevertheless, we do still see the pattern from the OTU analysis of $RAG^{-/-}$ changes being different from Black/6 and $MyD88^{-/-}$, again suggesting a role for the adaptive immune system in mediating these changes.

3.4.5 Correlation between microbiota functional content measured by whole genome sequencing, mouse genotype, and mycobacterial infection

In addition to the functional content predictions from PICRUSt, we used HUMAnN to make predictions about the gene content from our whole genome sequencing data [54]. Similar to previous analyses, there was a significant difference between genotypes (Figure 3-25). Only Balb/c with early timepoints removed and Black/6 samples were significantly different between *M. tuberculosis* infected and uninfected samples, although as with our other analyses, removing the early post-infection timepoints decreased the adjusted p-values for all genotypes (Figure 3-26 to Figure 3-29). However, very few KEGG orthologs were significantly associated with infection and there was no overlap in significant features between genotypes (Supplemental Table 3-S6, Figure 3-30). Furthermore, no significant KEGG orthologs from HUMAnN overlapped with significant orthologs from the PICRUSt analysis. Thus, although functional content can differentiate between *M. tuberculosis* infected and uninfected samples background, at least in

the Black/6 background, the overall genetic content seems to undergo fewer changes than at the OTU level.

3.4.6 Association between gut gene expression and *M. tuberculosis* infection

The presence of a gene in DNA does not give any indication of the expression of that gene, and thus although there were few changes at the DNA level, we also looked for changes in the RNA level. We used HUMAnN to predict the relative abundance of KEGG orthologs from our RNA-seq data, normalized the abundances using our HUMAnN whole genome sequencing data, and calculated fold change expression from pre-infection. Supplemental Figure 3-S1 shows all KEGG orthologs identified as significantly changed from day -3 in at least one timepoint (data also listed in Supplemental Table 3-S7). As in the PICRUSt analysis, there was very little overlap between mouse genotypes (Figure 3-31). One KEGG ortholog was shared by the knockout mice, 12 by Black/6 and MyD88$^{-/-}$, and two between the wild-type strains. However, like at the DNA level, Black/6 had the most orthologs identified as significant (Supplemental Figure 3-S1b), and the most overlap was between Black/6 and MyD88$^{-/-}$, suggesting that the response of MyD88$^{-/-}$ microbiota to infection is more like wild-type than RAG$^{-/-}$ microbiota, an observation echoed at the OTU level.

3.4.7 The changes in gut microbial composition and gene content are specific to *M. tuberculosis*

One important question is whether the changes that we observed were specific to *M. tuberculosis* or were the result of a more generalized response to infection. To address this question, we used wild-type mice infected with other mycobacterial species (Table 3-1). One

control group we used was Black/6 mice infected by *M. smegmatis*. Unfortunately, the *M. smegmatis* cage was the only one with a potential caging effect (Figure 3-5b), which can be seen in the relative abundances even in pre-infection for these mice (Figure 3-4). As a result, *M. smegmatis* infected samples were significantly different from uninfected samples when comparing the relative abundances of OTUs from 16S rDNA sequencing after removing early timepoints (Figure 3-32). However, *M. smegmatis* infected and uninfected samples were not significantly different in any other analysis (Figure 3-32 and 3-33). Thus, although there may have been a caging effect, it was not enough to differentiate these samples from the uninfected mice in most analyses. In contrast, *M. smegmatis*-infected samples were significantly different from *M. tuberculosis*-infected samples in all analyses (Figure 3-34 and 3-35). Thus, the microbiota seems to be responding differently to *M. tuberculosis* infection than to *M. smegmatis* infection, suggesting the changes we saw were not due to a generalized activation of the immune system.

The other group we analyzed was Balb/c mice infected with *M. avium.* Although *M. avium* was not cleared from the mouse (Figure 3-1a), there was no significant difference between *M. avium* infected and uninfected sample in the 16S rDNA analyses, although whole genome sequencing (both OTU and functional content) was significant (Figure 3-36 and 3-37). In contrast, in all analyses there was a highly significant difference between *M. tuberculosis* and *M. avium* infected samples (Figure 3-38 and 3-39), confirming the findings with *M. smegmatis* and suggesting that the differences we described above were specific to *M. tuberculosis.*

3.5 Discussion

Our previous work showed that there was a significant change in the gut microbiota of Balb/c mice in response to *M. tuberculosis* infection (Chapter 2; [53]). Here we have confirmed

those results and showed that despite an altered microbial composition in uninfected mice as a result of changes in experimental design, there was a significant change in the gut microbiota of Balb/c as measured by 16S rDNA sequencing by day 10 (Figure 3-7). Furthermore, these changes also occurred in Black/6 and MyD88$^{-/-}$ mice, showing that these changes are not specific to Balb/c mice (Figures 3-8 and 3-9). In contrast, there was not a significant difference in RAG$^{-/-}$ mice, indicating that the adaptive immune system plays a role in mediating these changes (Figure 3-10). Interestingly, the changes in the other mouse genotypes are occurring before the adaptive immune system begins to control the infection (Figures 3-1 and 3-7 through 3-9), suggesting that the baseline microbiota also plays role. These observations were echoed in our whole genome sequencing data, despite a paucity of detected OTUs (Figures 3-14 through 3-18).

From our whole genome sequencing data, *Enterorhabdus caecimuris* was identified as being significantly associated with *M. tuberculosis* CFU counts in both the lungs and spleens. *E. caecimuris* are gram positive non-spore-forming rods that were isolated from a mouse model of intestinal inflammation and thus may have some interaction with the immune system [63]. An OTU identified from 16S rDNA sequencing as being associated with *M. tuberculosis* CFU counts was in the genera *Adlercreutzia*. This genera has also been identified in humans, and was significantly associated with infection in our previous studies, and in all mouse genotypes in this study except RAG$^{-/-}$ (which does not undergo microbiota changes with infection) [64,65]. Thus, this OTU may be a potential biomarker of *M. tuberculosis* infection in patients with an intact adaptive immune system.

In addition to changes at the OTU level, this data enabled us to look at changes in gene content and expression. Only Black/6 mice showed significant changes in response to *M. tuberculosis* infection in both 16S rDNA and whole genome sequencing data, but MyD88$^{-/-}$ and

RAG$^{-/-}$ mice were also significantly changed in 16S rDNA sequencing (Figure 3-19 through 3-30). Interestingly, in the PICRUSt predictions and RNA-seq data MyD88$^{-/-}$ and Black/6 mice shared the most significantly changed KEGG orthologs, supporting the OTU results that the changes in response to infection are mediated by the adaptive immune system (Figures 3-24 and 3-31). These significant KEGG orthologs spanned a range of functions, including dehydrogenases, DNA polymerase, and ribosomal proteins, suggesting that there is not a specific pathway that is altered.

Taken together, both the 16S rDNA sequencing and whole genome sequencing data indicate that the microbiome of RAG$^{-/-}$ mice responds differently to *M. tuberculosis* infection compared to the other genotypes, both in OTU composition and gene content. There is growing evidence of a strong association between the microbiota and the adaptive immune system, and this interaction may play a role in *M. tuberculosis* infection [13,46,60]. The fact that the changes we observed are mediated by the adaptive immune system is of particular concern in patients with a compromised adaptive immune system, such as patients infected with the Human Immunodeficiency Virus (HIV). *M. tuberculosis* is a leading killer of HIV patients, and there is evidence that the microbiota plays a role in susceptibility to HIV [23,66-69]. Furthermore, the microbiota is disrupted by HIV infection [70-73]. Thus, given our findings, more work needs to be done on elucidating the interaction between the microbiota, HIV and other disease that suppress the immune system, and *M. tuberculosis.*

Our analysis of *M. smegmatis* and *M. avium* infected mice revealed that the microbiota does not change at the OTU or gene content level in response to aerosol infection with these mycobacteria, indicating that the changes we detected are specific to *M. tuberculosis* and are not just the result of an activation of the immune system (Figure 3-32 through 3-39). *M.*

tuberculosis induces a unique immune response compared to many other pathogens, and given the fact that the immune system is mediating the microbiota changes, as evidenced by that lack of change in OTUs in RAG$^{-/-}$ mice, these changes may be useful as biomarkers of *M. tuberculosis* infection [12]. Thus, here we have shown that the gut microbiota specifically responds to aerosol *M. tuberculosis* infection by day 10 post-infection at both the OTU and gene content levels. These changes are less pronounced in RAG$^{-/-}$ mice, especially in terms of changes in OTU composition, suggesting that the adaptive immune system plays a role in mediating this interaction between *M. tuberculosis* and the gut microbiota.

3.6 Figures

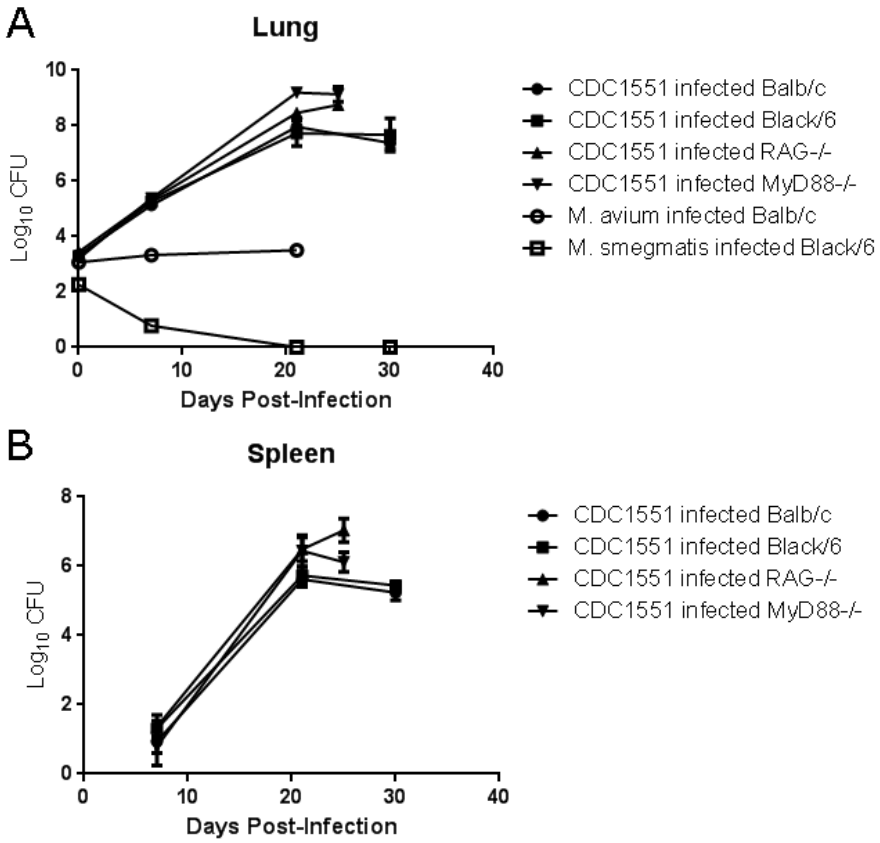
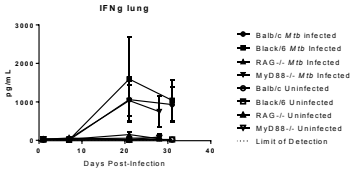
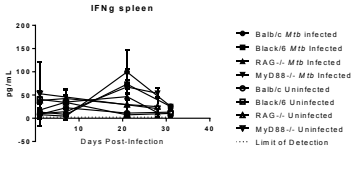
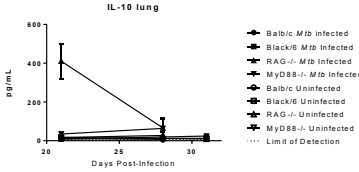


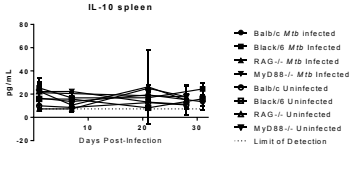
Figure 3-1. Bacterial burden in lungs and spleen.

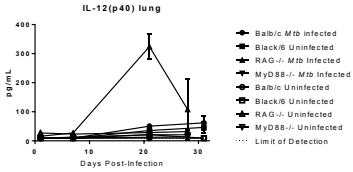
Colony forming units (CFU) in A) lungs and B) spleen.

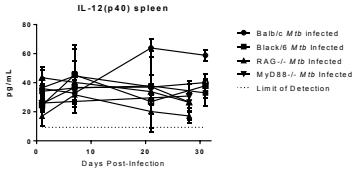

IFNg lung
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

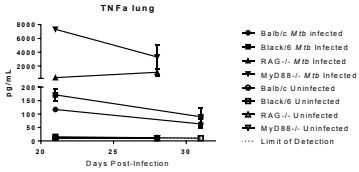

IFNg spleen
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

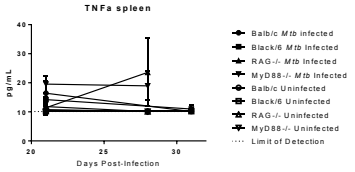

IL-10 lung
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

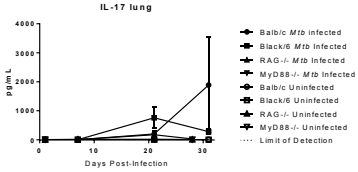

IL-10 spleen
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

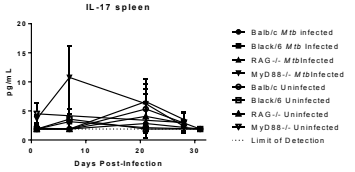

IL-12(p40) lung
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Uninfected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

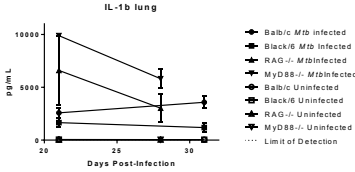

IL-12(p40) spleen
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Limit of Detection

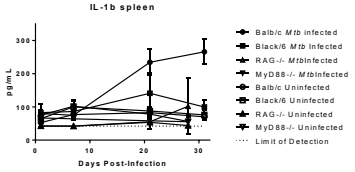

TNFa lung
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

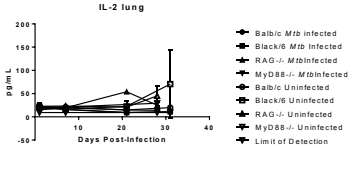

TNFa spleen
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

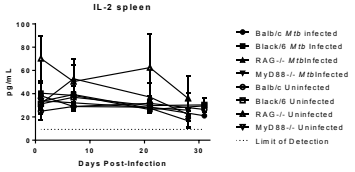

IL-17 lung
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

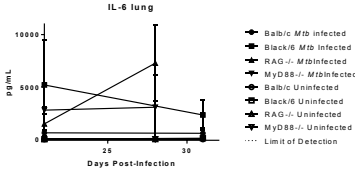

IL-17 spleen
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection


IL-1b lung
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection


IL-1b spleen
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection


IL-2 lung
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection


IL-2 spleen
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection


IL-6 lung
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

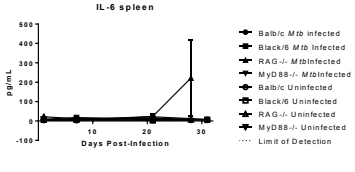

IL-6 spleen
pg/mL
Days Post-Infection
Balb/c Mtb infected
Black/6 Mtb Infected
RAG-/- Mtb Infected
MyD88-/- Mtb Infected
Balb/c Uninfected
Black/6 Uninfected
RAG-/- Uninfected
MyD88-/- Uninfected
Limit of Detection

Figure 3-2. Lung and spleen cytokine levels.

Levels of each cytokine in each mouse genotype, *M. tuberculosis (Mtb)* infected or uninfected. The limit of detection for each cytokine is indicated with a dotted line.

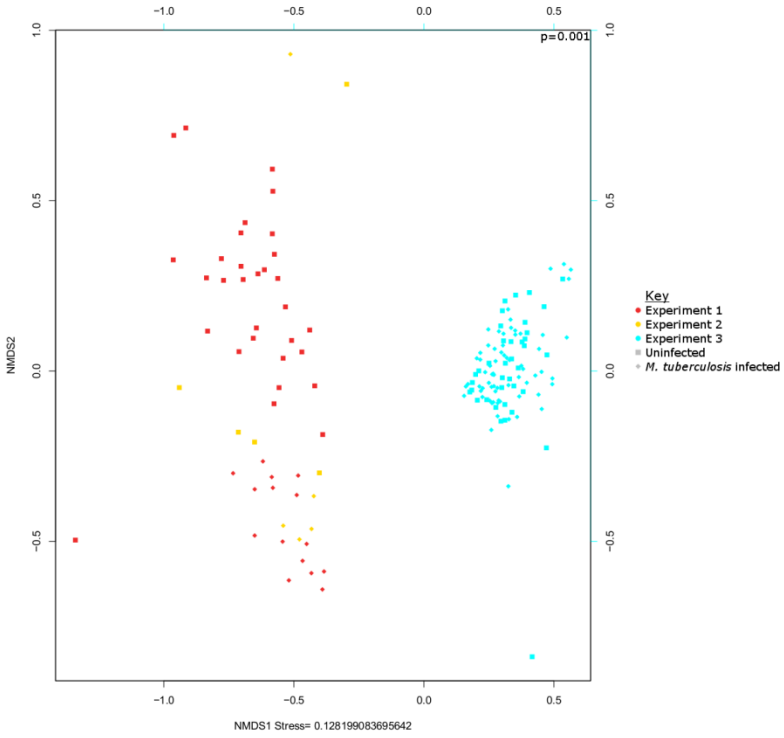


Figure 3-3. NMDS plot of all Balb/c uninfected or *M. tuberculosis* infected samples from all experiments.

Points are colored by experiment and shaped by infection status. Experiment 1 was a longitudinal study following 5 mice from pre-infection to death. Experiment 2 was a single timepoint comparing 5 infected mice to 5 age-matched uninfected mice. (Published in [53]). Experiment 3 was the new experiment following different mouse genotypes for one month (described in this chapter). The P-value was calculated for experiment using adonis in Vegan.

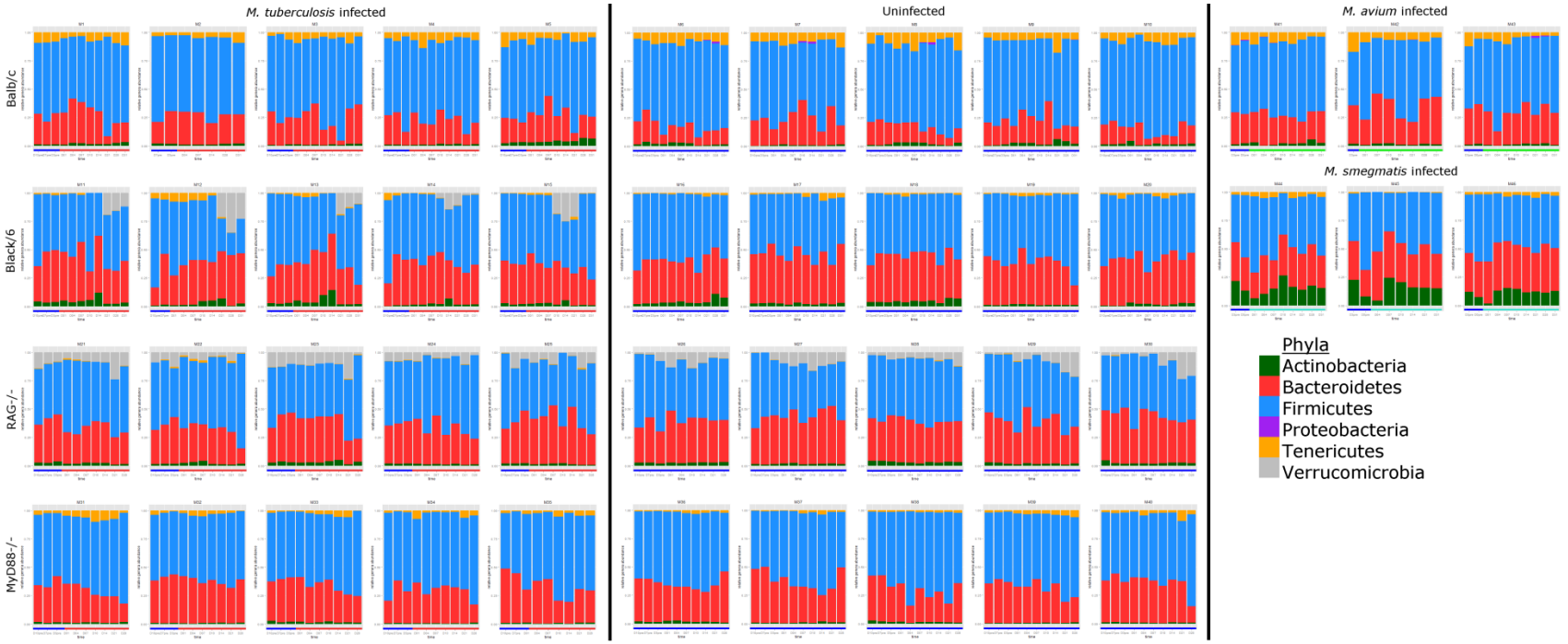

M. tuberculosis infected
Uninfected
M. avium infected
M. smegmatis infected
Balb/c
Black/6
RAG-/-
MyD88-/-
Phyla
Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Tenericutes
Verrucomicrobia

Figure 3-4. Relative abundance of OTUs identified by QIIME from 16S sequencing.

Each bar represents a different timepoint. Each graph is a different mouse. A blue line below a bar indicates the sample was uninfected, while a red line indicates *M. tuberculosis* infected, a green line indicates *M. avium* infection and a turquoise line indicates *M. smegmatis* infected. The relative abundance colors are based on phyla.

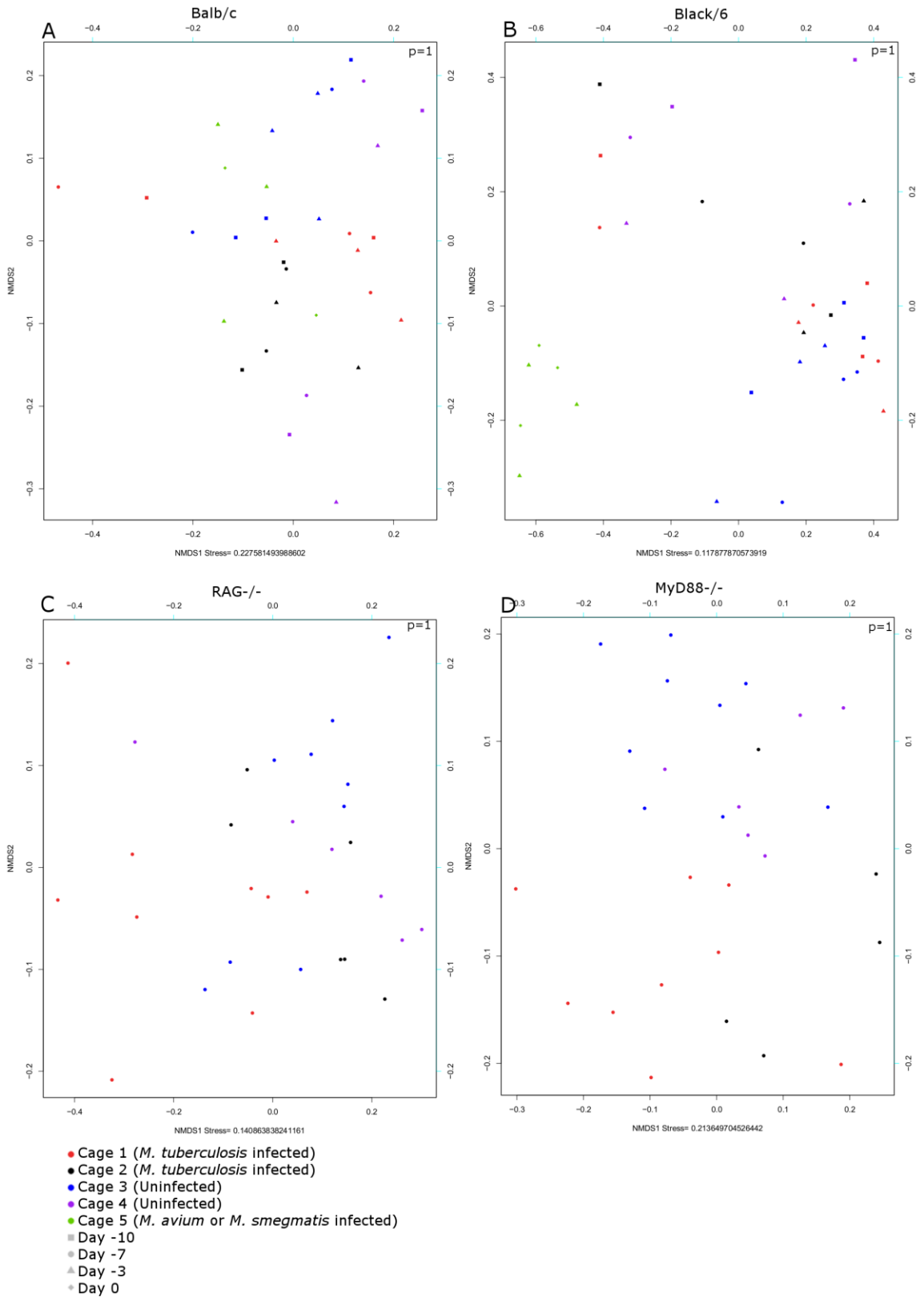


Figure 3-5. No significant caging effect was detected in this experiment.

NMDS plots of OTUs predicted by QIIME from 16S sequencing data in pre-infection timepoints for (A) Balb/c (B) Black/6 (C) RAG $^{-/-}$ and (D) MyD88$^{-/-}$. Points are colored by cage and shaped by timepoint. P values are for cage number as calculated by adonis in Vegan, stratified by mouse number.

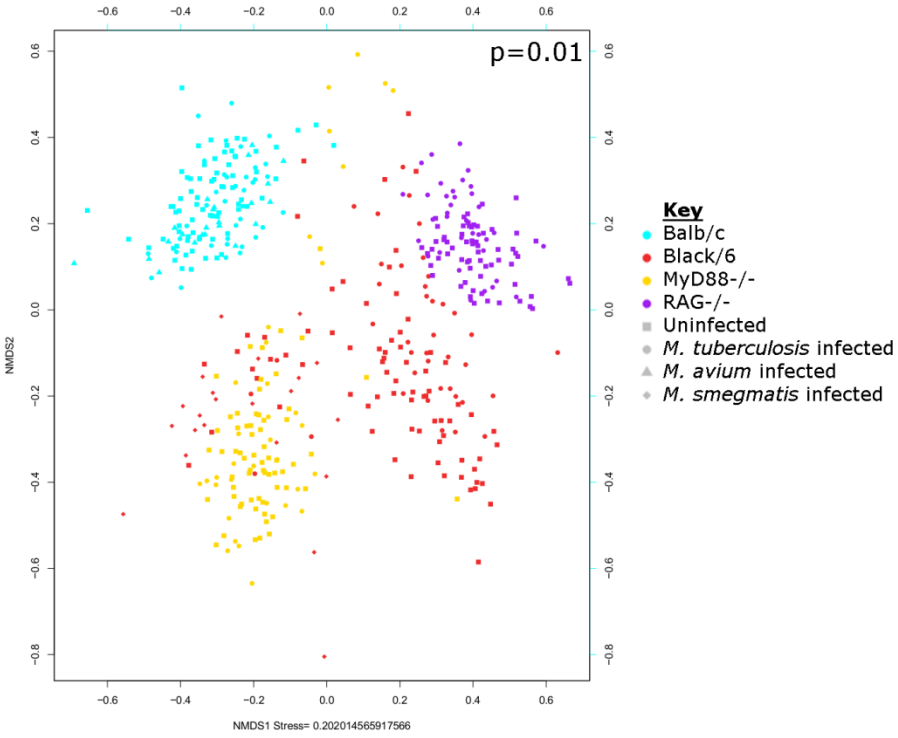


Figure 3-6. Gut microbial composition as assessed by QIIME from 16S sequencing is significantly different between mouse genotypes.

All samples from 16S sequencing, colored by genotype and shaped by infection status. P value is for genotype, as calculated by adonis in Vegan and stratified by infecting organism.

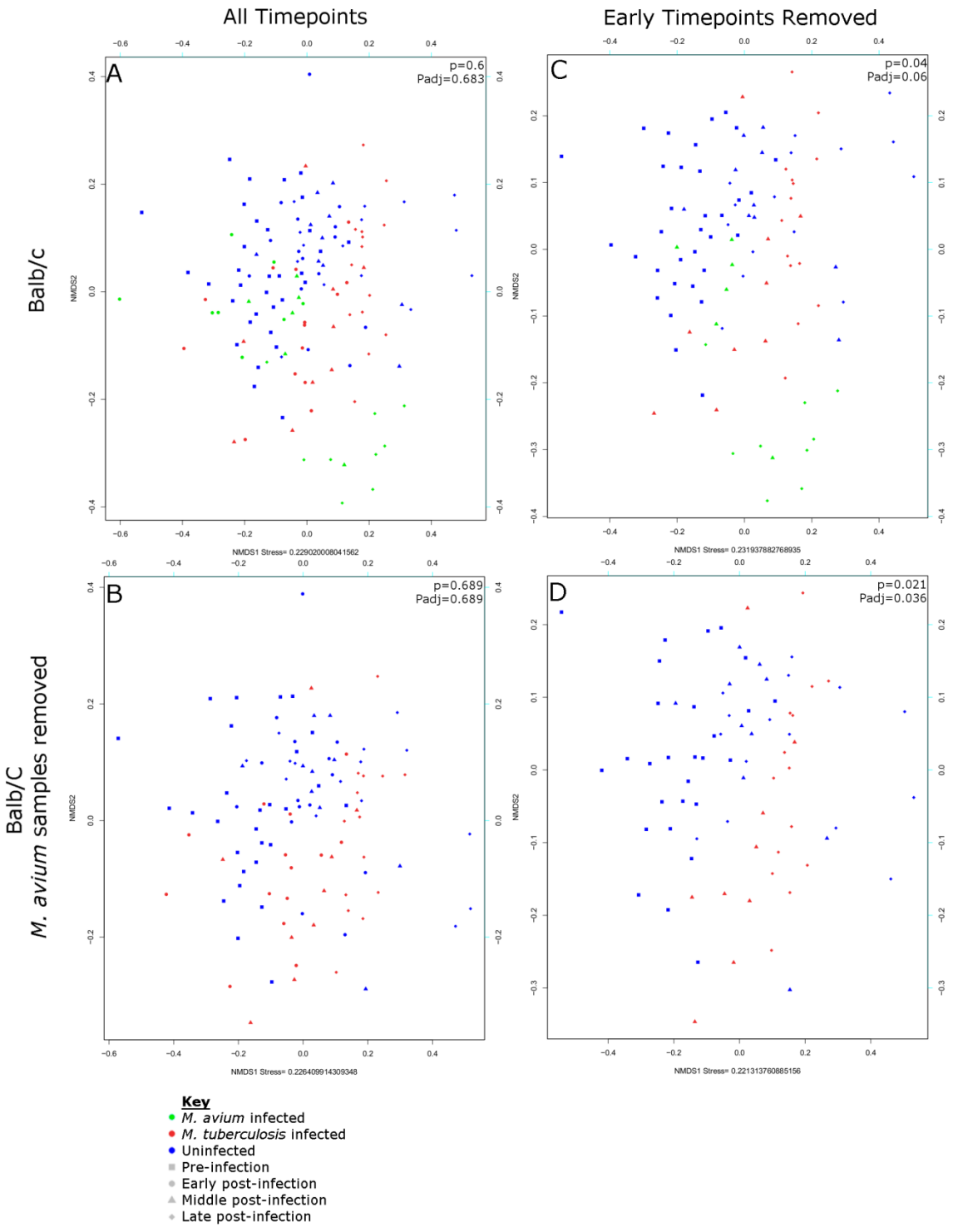


Figure 3-7. Gut microbial composition of Balb/c mice, as predicted by QIIME from 16S sequencing, changes by day 10 post-infection in response to *M. tuberculosis* infection.

Balb/c 16S sequencing samples colored by infection status and shaped by timepoint. A) All Balb/c samples. B) Balb/c samples with *M. avium* infected mice removed. C) Balb/c samples with early post-infection timepoints removed. D) Balb/c samples with early post-infection samples and *M. avium* infected mice removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

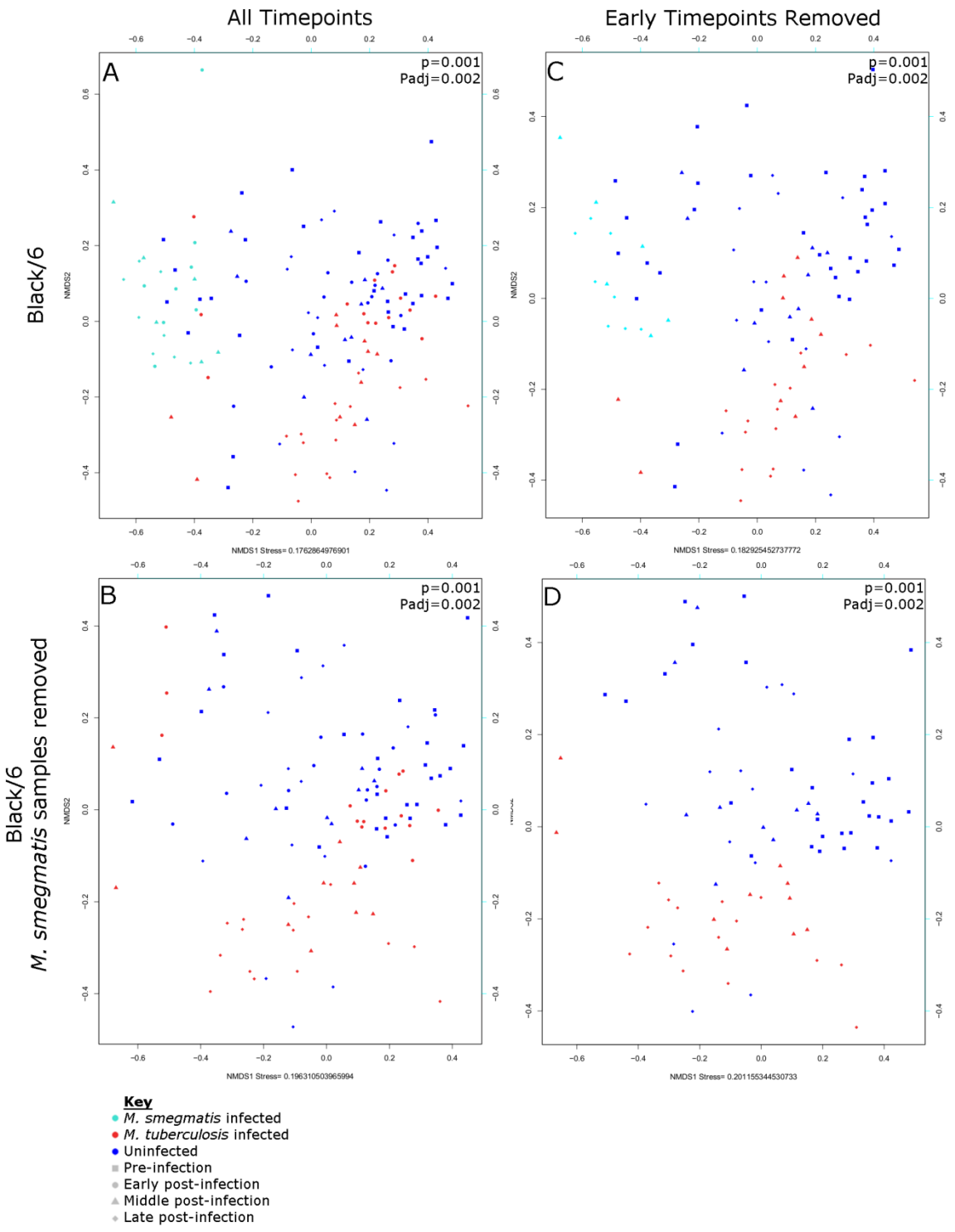


Figure 3-8. Gut microbial composition of Black/6 mice, as predicted by QIIME from 16S sequencing, changes in response to *M. tuberculosis* infection.

Black/6 16S sequencing samples colored by infection status and shaped by timepoint. A) All Black/6 samples. B) Black/6 samples with *M. smegmatis* infected mice removed. C) Black/6 samples with early post-infection timepoints removed. D) Black/6 samples with early post-infection samples and *M. smegmatis* infected mice removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

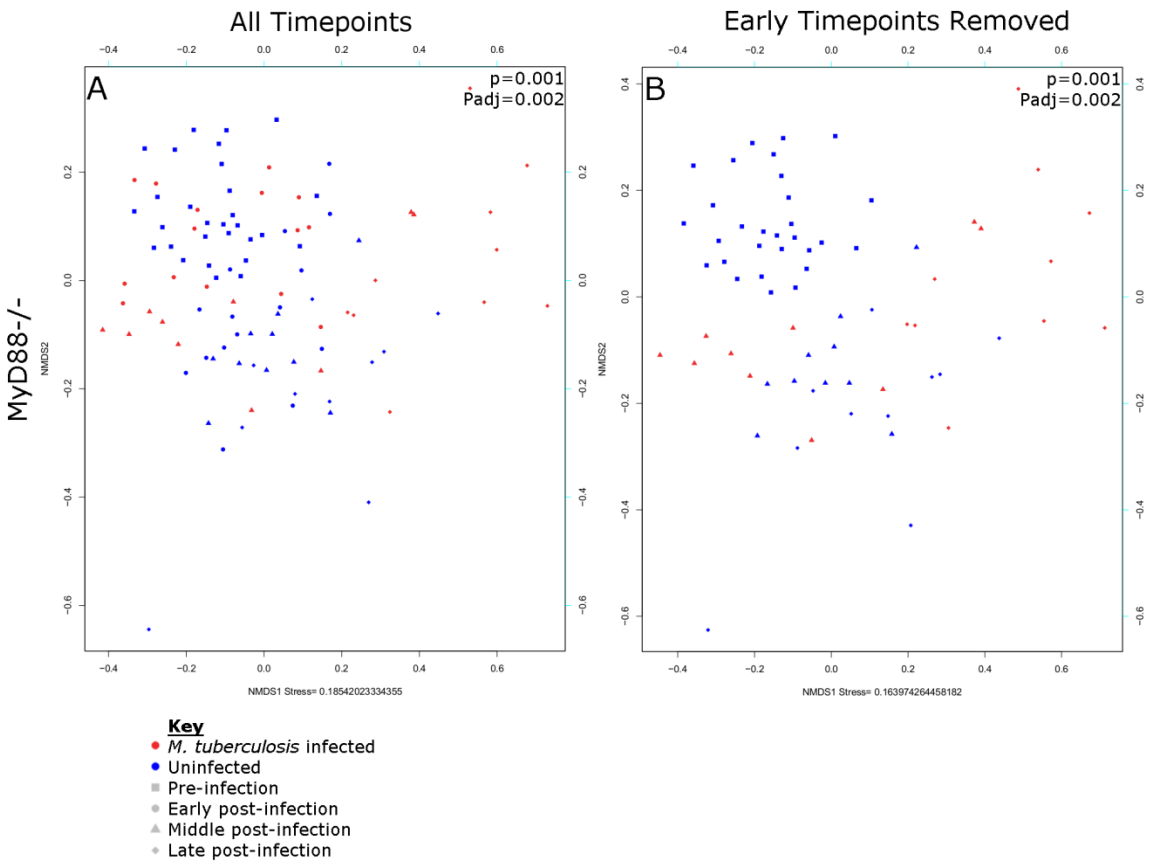


Figure 3-9. Gut microbial composition of MyD88$^{-/-}$ mice, as predicted by QIIME from 16S sequencing, changes in response to *M. tuberculosis* infection.

MyD88$^{-/-}$ 16S sequencing samples colored by infection status and shaped by timepoint. A) All MyD88$^{-/-}$samples. B) MyD88$^{-/-}$samples with early post-infection timepoints removed. P values

are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

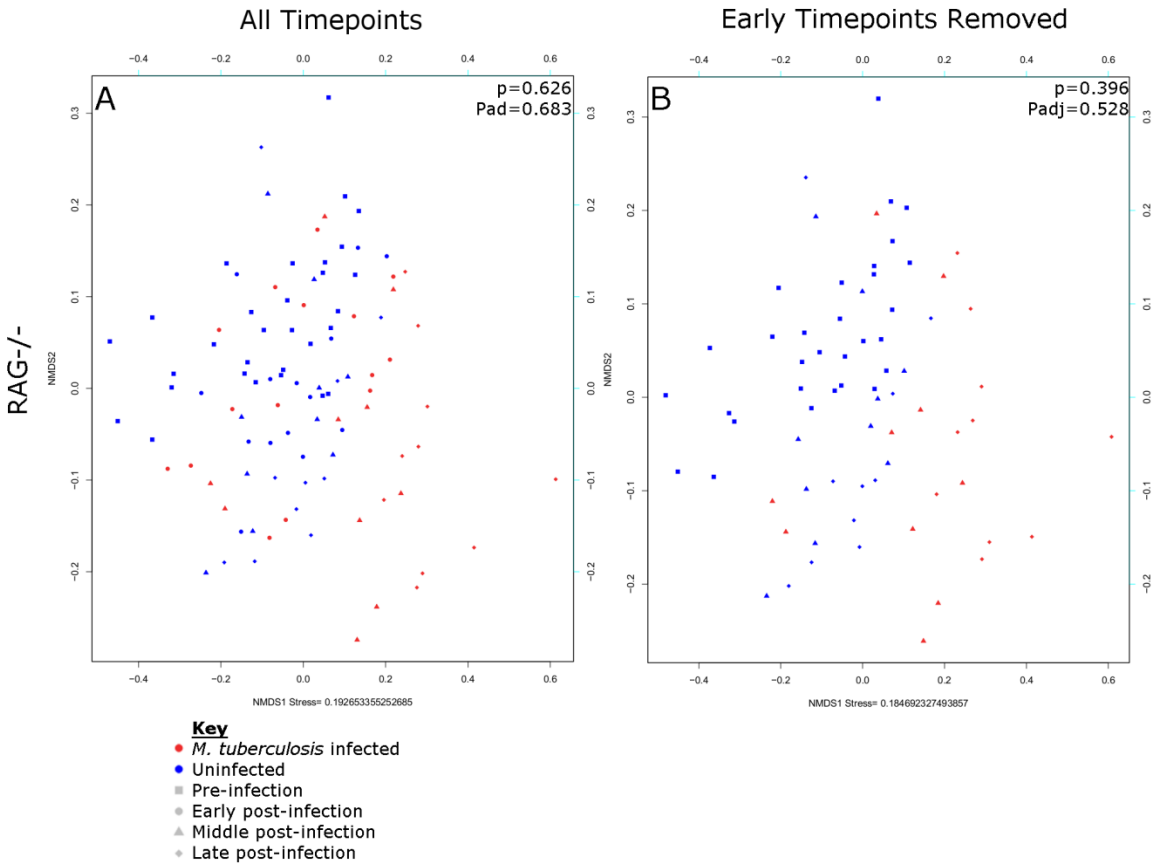
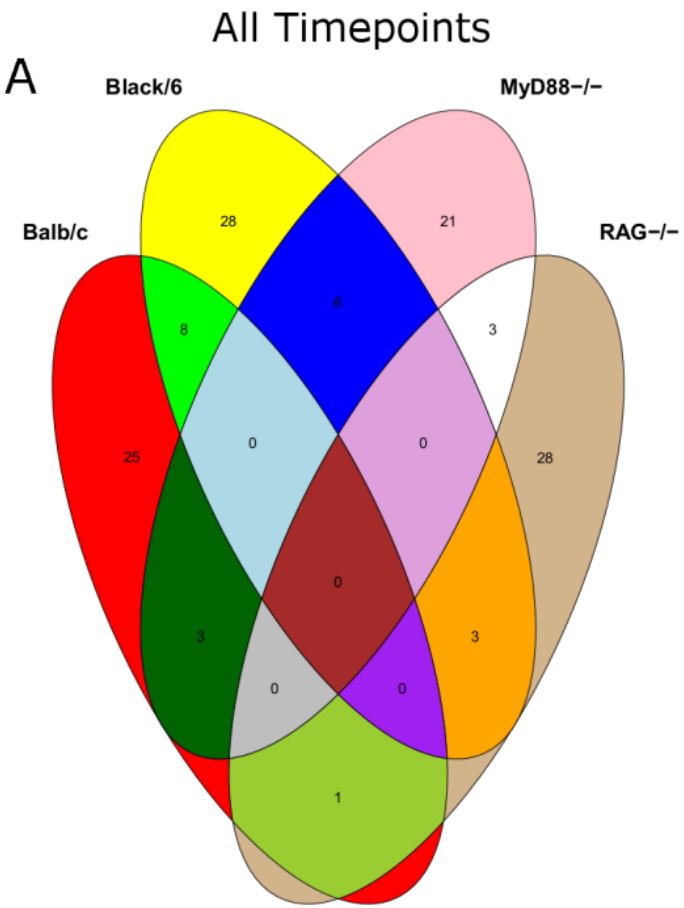


Figure 3-10. Gut microbial composition of RAG$^{-/-}$ mice, as predicted by QIIME from 16S sequencing, does not change in response to *M. tuberculosis* infection.

RAG$^{-/-}$ 16S sequencing samples colored by infection status and shaped by timepoint. A) All RAG$^{-/-}$ samples. B) RAG$^{-/-}$samples with early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

Significant OTUs



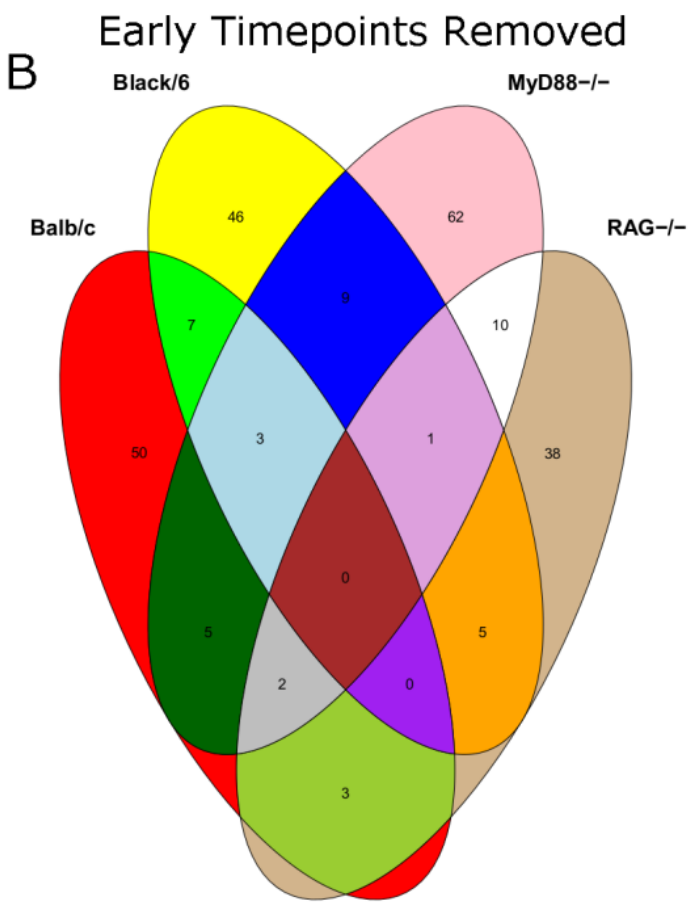


Figure 3-11. Overlap between genotypes in OTUs from 16S sequencing significantly different between *M. tuberculosis* infected and uninfected samples.

Venn diagram of overlaps of significant OTUs for each genotype with all timepoints (A) or with early timepoints removed (B).

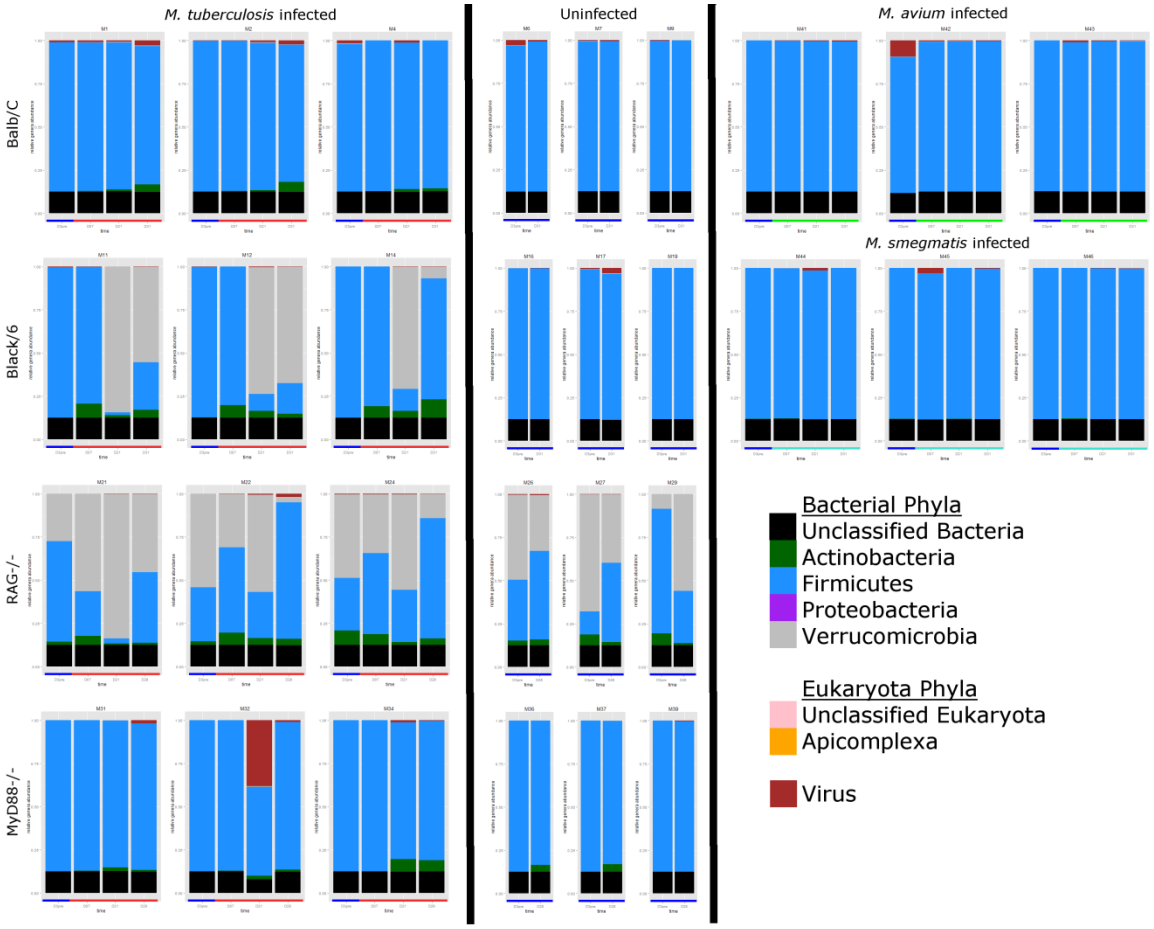


Figure 3-12. Relative abundance of OTUs identified by MetaPhlAn from whole genome sequencing data.

Each bar represents a different timepoint. Each graph represents a different mouse. A blue line below a bar indicates the sample was uninfected, while a red line indicates *M. tuberculosis*

infected, a green line indicates *M. avium* infection and a turquoise line indicates *M. smegmatis* infected. The relative abundance colors are based on phyla.

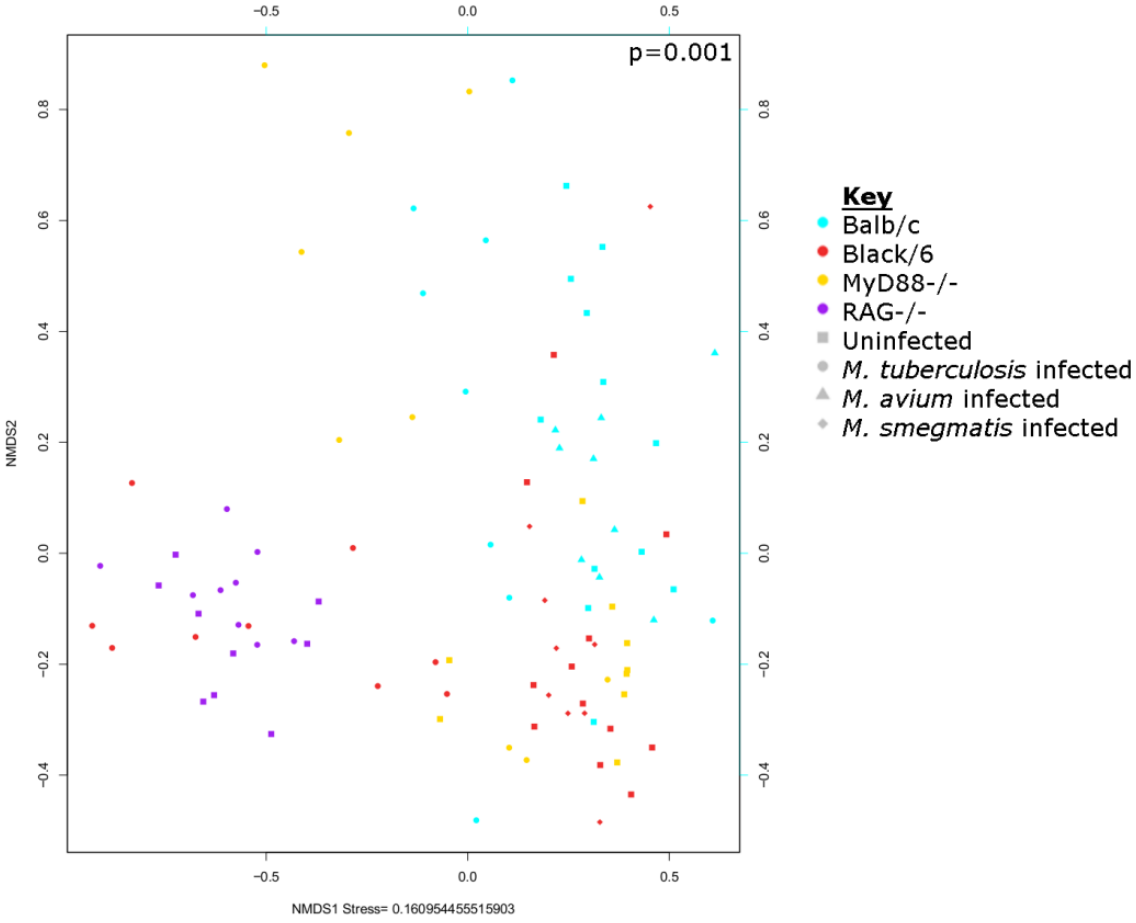


Figure 3-13. Gut microbial composition as predicted by MetaPhlAn from whole genome sequencing is significantly different between mouse genotypes.

All samples from 16S sequencing, colored by genotype and shaped by infection status. P value is for genotype, as calculated by adonis in Vegan and stratified by infecting organism.

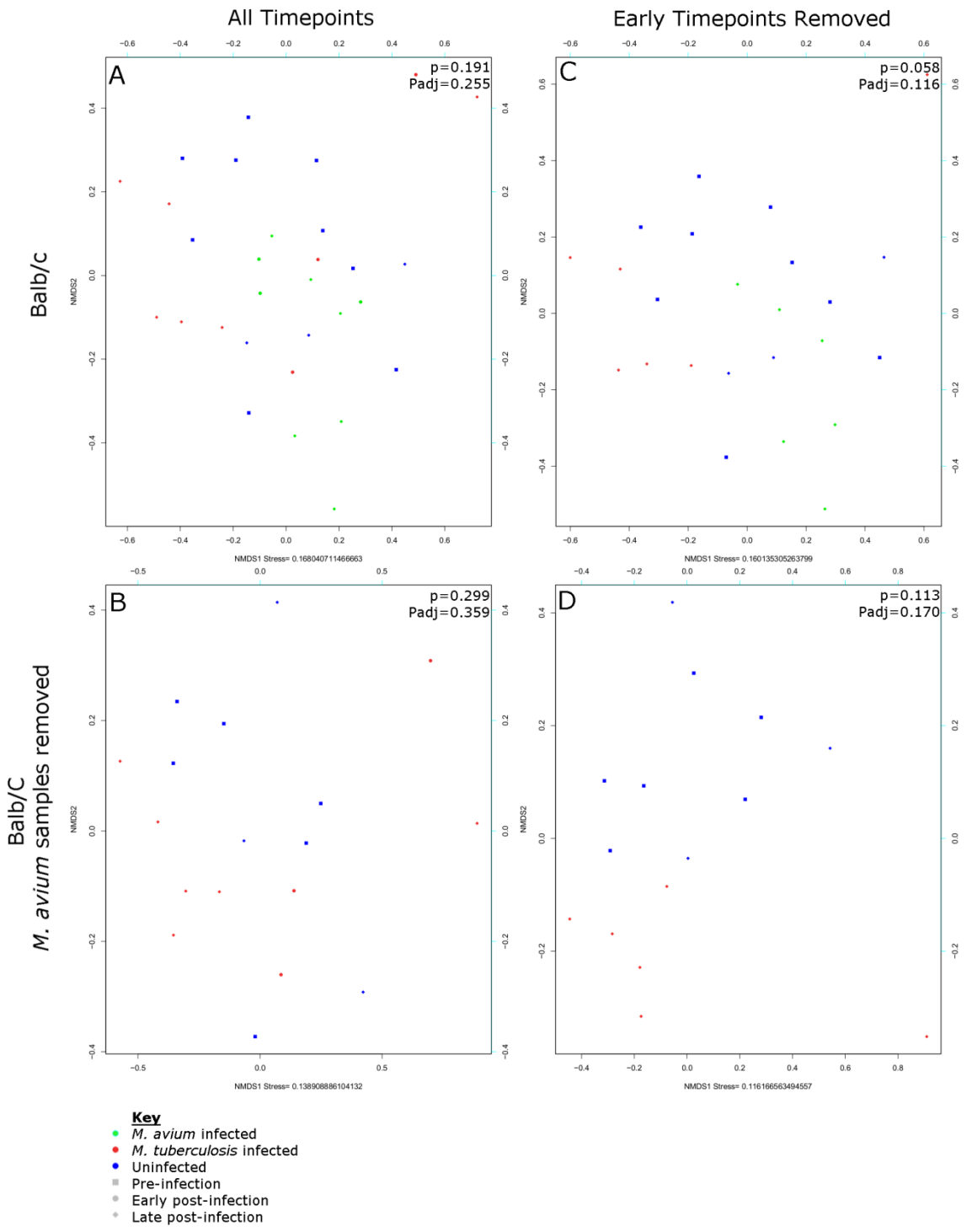


Figure 3-14. Gut microbial composition of Balb/c mice, as predicted by MetaPhlAn from whole genome sequencing, does not respond to mycobacterial infection.

Balb/c whole genome sequencing samples colored by infection status and shaped by timepoint. A) All Balb/c samples. B) Balb/c samples with *M. avium* infected mice removed. C) Balb/c samples with early post-infection timepoints removed. D) Balb/c samples with early post-infection samples and *M. avium* infected mice removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

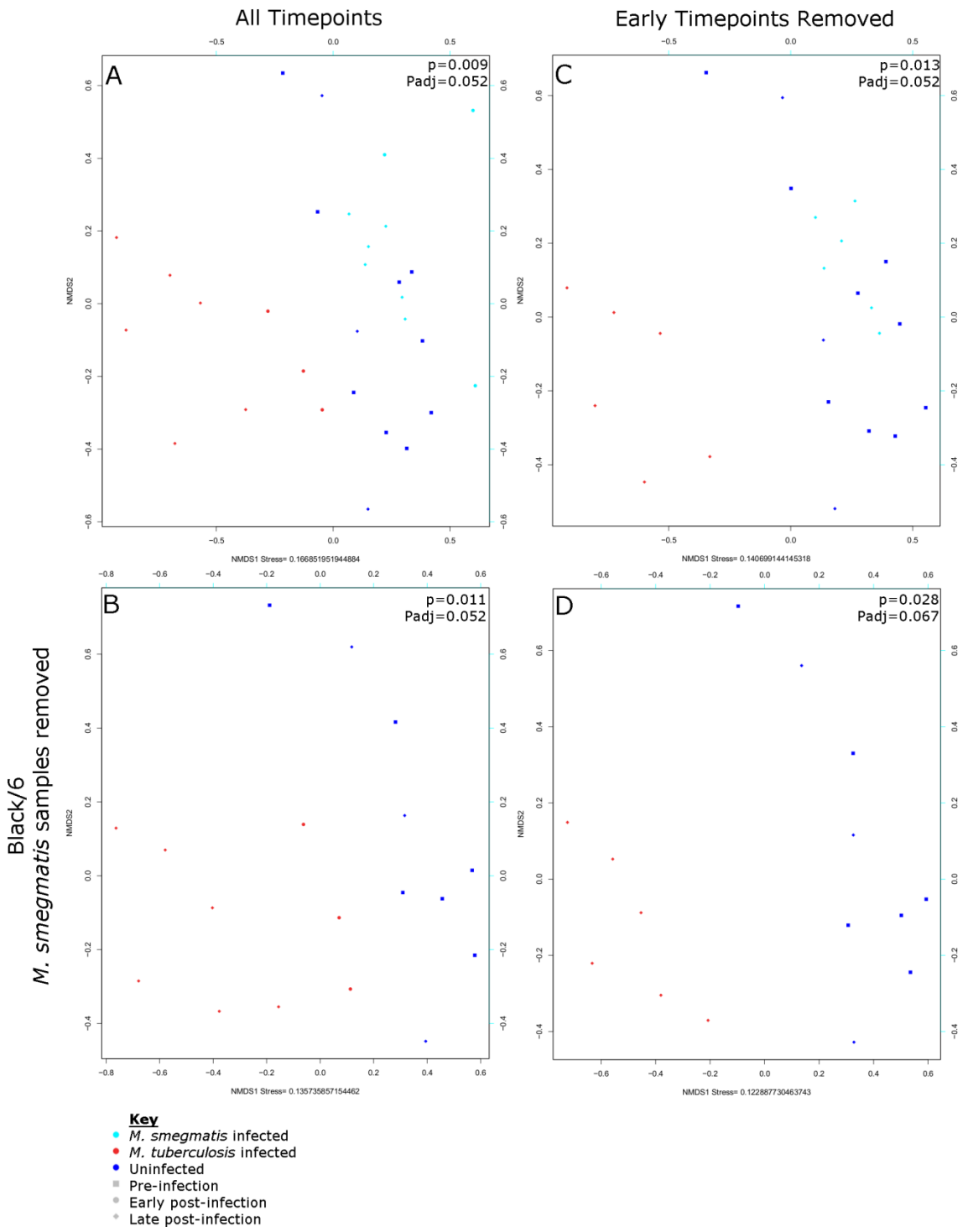


Figure 3-15. Gut microbial composition of Black/6 mice, as predicted by MetaPhlAn from whole genome sequencing, does not respond to mycobacterial infection.

Black/6 whole genome sequencing samples colored by infection status and shaped by timepoint. A) All Black/6 samples. B) Black/6 samples with *M. smegmatis* infected mice removed. C) Black/6 samples with early post-infection timepoints removed. D) Black/6 samples with early post-infection samples and *M. smegmatis* infected mice removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

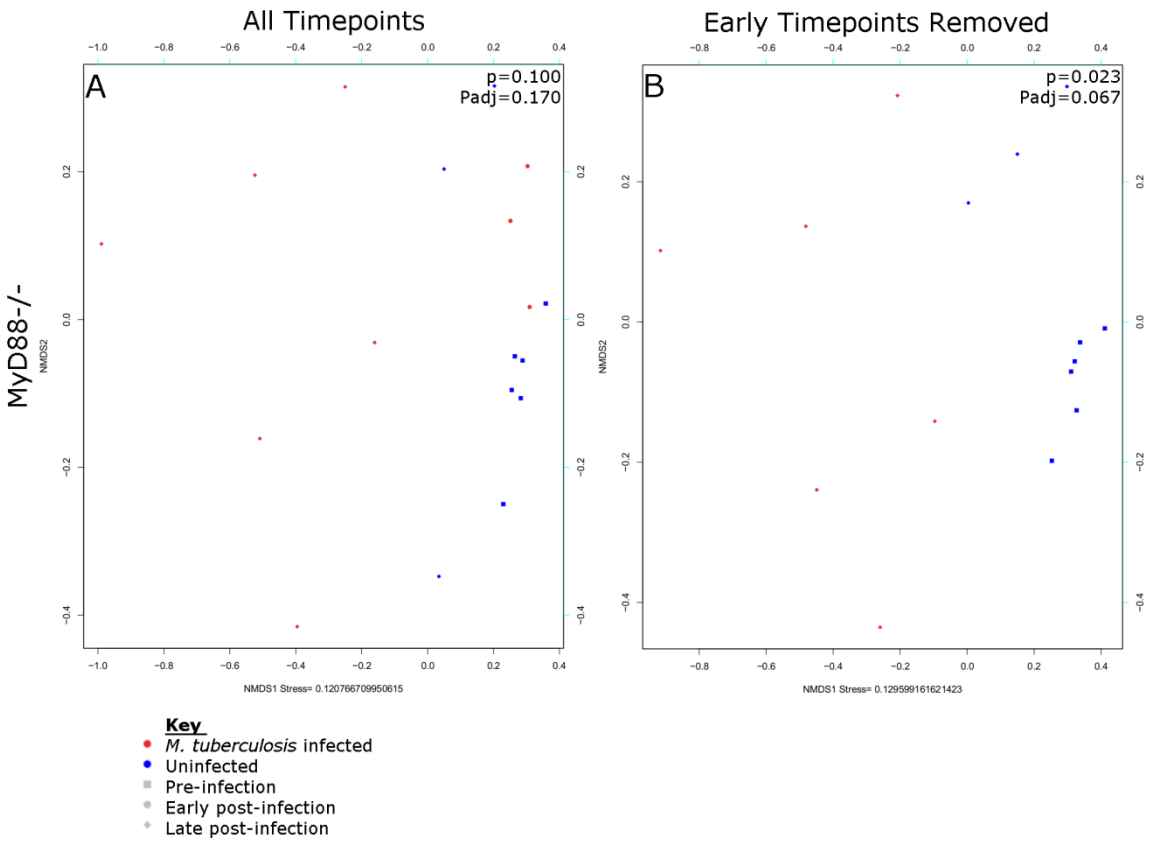


Figure 3-16. Gut microbial composition of MyD88$^{-/-}$ mice, as predicted by MetaPhlan from whole genome sequencing, does not respond to *M. tuberculosis* infection.

MyD88$^{-/-}$whole genome sequencing samples colored by infection status and shaped by timepoint. A) All MyD88$^{-/-}$ samples. B) MyD88$^{-/-}$ samples with early post-infection timepoints

removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

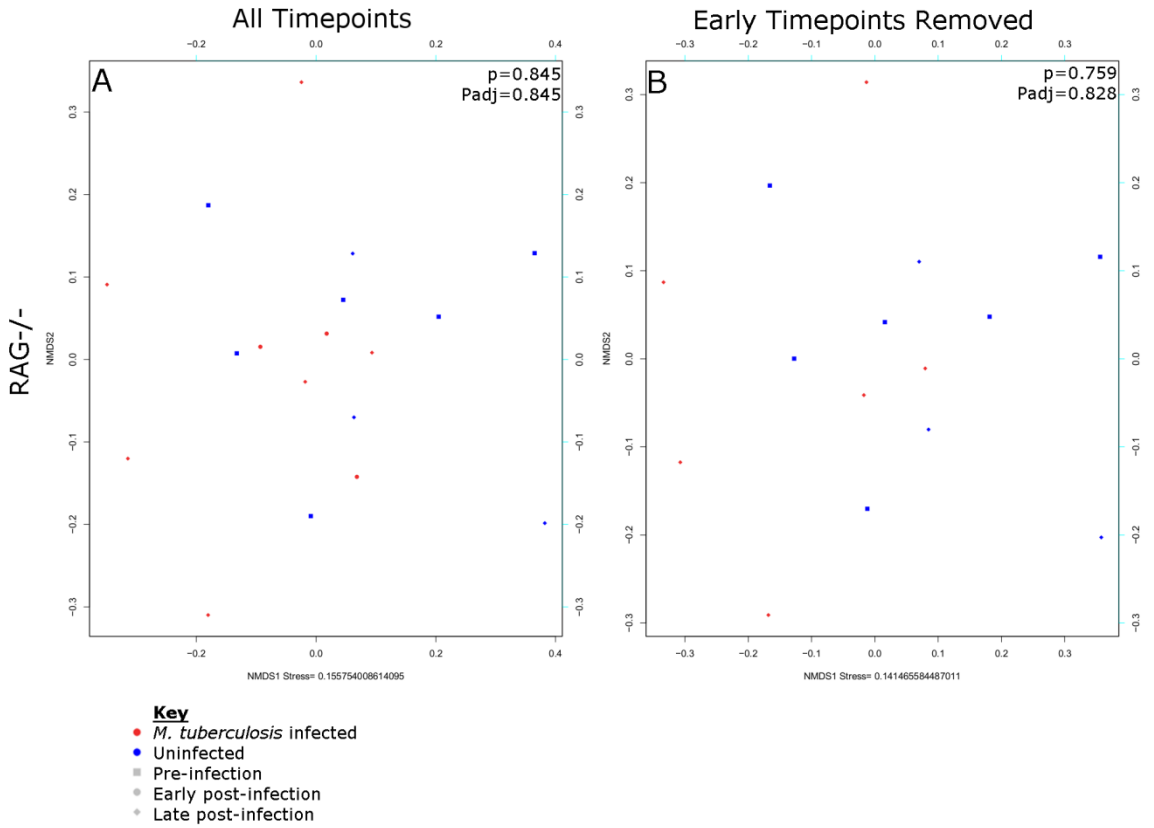


Figure 3-17. Gut microbial composition of RAG$^{-/-}$ mice, as predicted by MetaPhlAn from whole genome sequencing, does not respond to *M. tuberculosis* infection.

RAG$^{-/-}$whole genome sequencing samples colored by infection status and shaped by timepoint. A) All RAG$^{-/-}$ samples. B) RAG$^{-/-}$ samples with early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

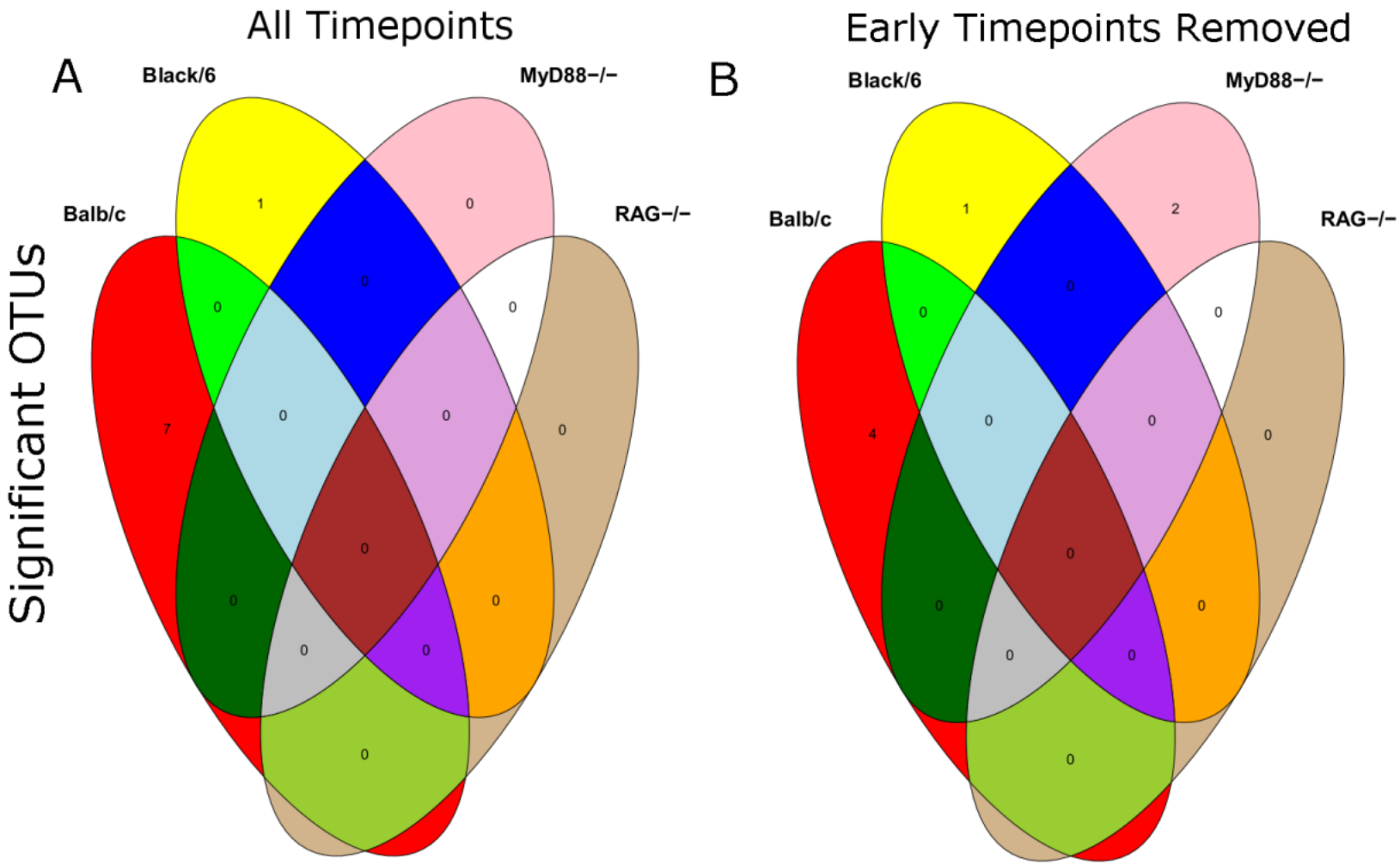

All Timepoints
A
Black/6
MyD88−/−
Balb/c
RAG−/−
Significant OTUs
1
0
0
0
0
7
0
0
0
0
0
0
0
0
0
Early Timepoints Removed
B
Black/6
MyD88−/−
Balb/c
RAG−/−
1
2
0
0
0
4
0
0
0
0
0
0
0
0
0

Figure 3-18. Overlap between genotypes in OTUs from whole genome sequencing significantly different between *M. tuberculosis* infected and uninfected samples.

Venn diagram of overlaps of significant OTUs for each genotype with all timepoints (A) or with early timepoints removed (B).

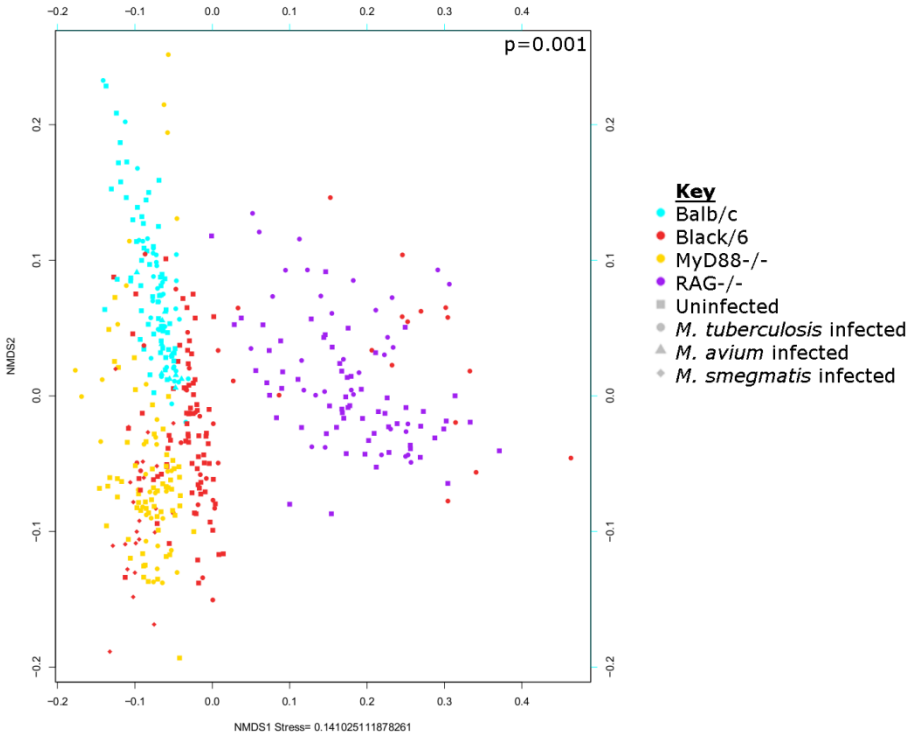


Figure 3-19. Gut gene content as predicted by PICRUSt from 16S sequencing is significantly different between mouse genotypes.

All samples from 16S sequencing, colored by genotype and shaped by infection status. P value is for genotype, as calculated by adonis in Vegan and stratified by infecting organism.

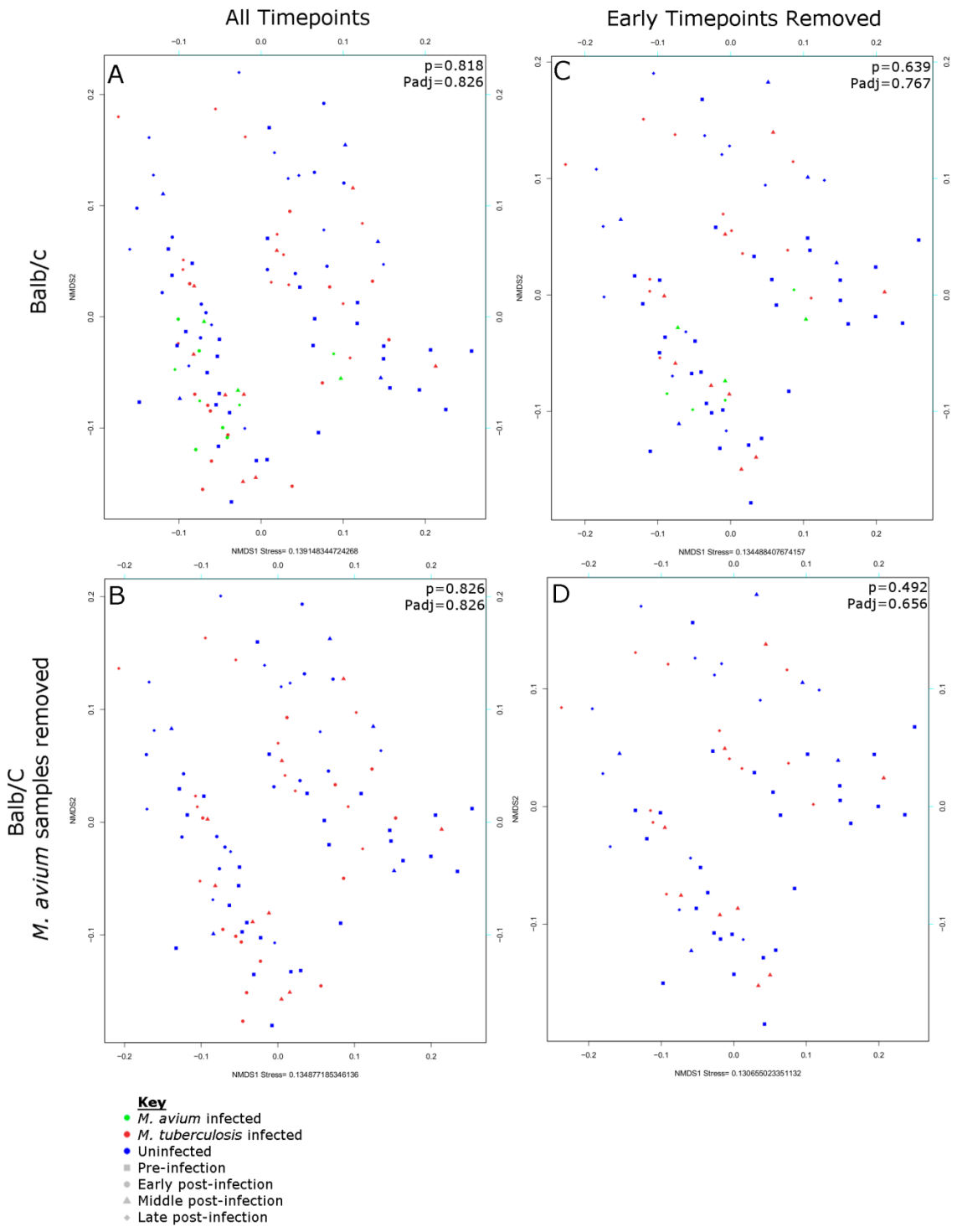


Figure 3-20. Gut gene content of Balb/c mice, as predicted by PICRUSt from 16S sequencing, does not respond to mycobacterial infection.

Balb/c 16S sequencing samples colored by infection status and shaped by timepoint. A) All Balb/c samples. B) Balb/c samples with *M. avium* infected mice removed. C) Balb/c samples with early post-infection timepoints removed. D) Balb/c samples with early post-infection samples and *M. avium* infected mice removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

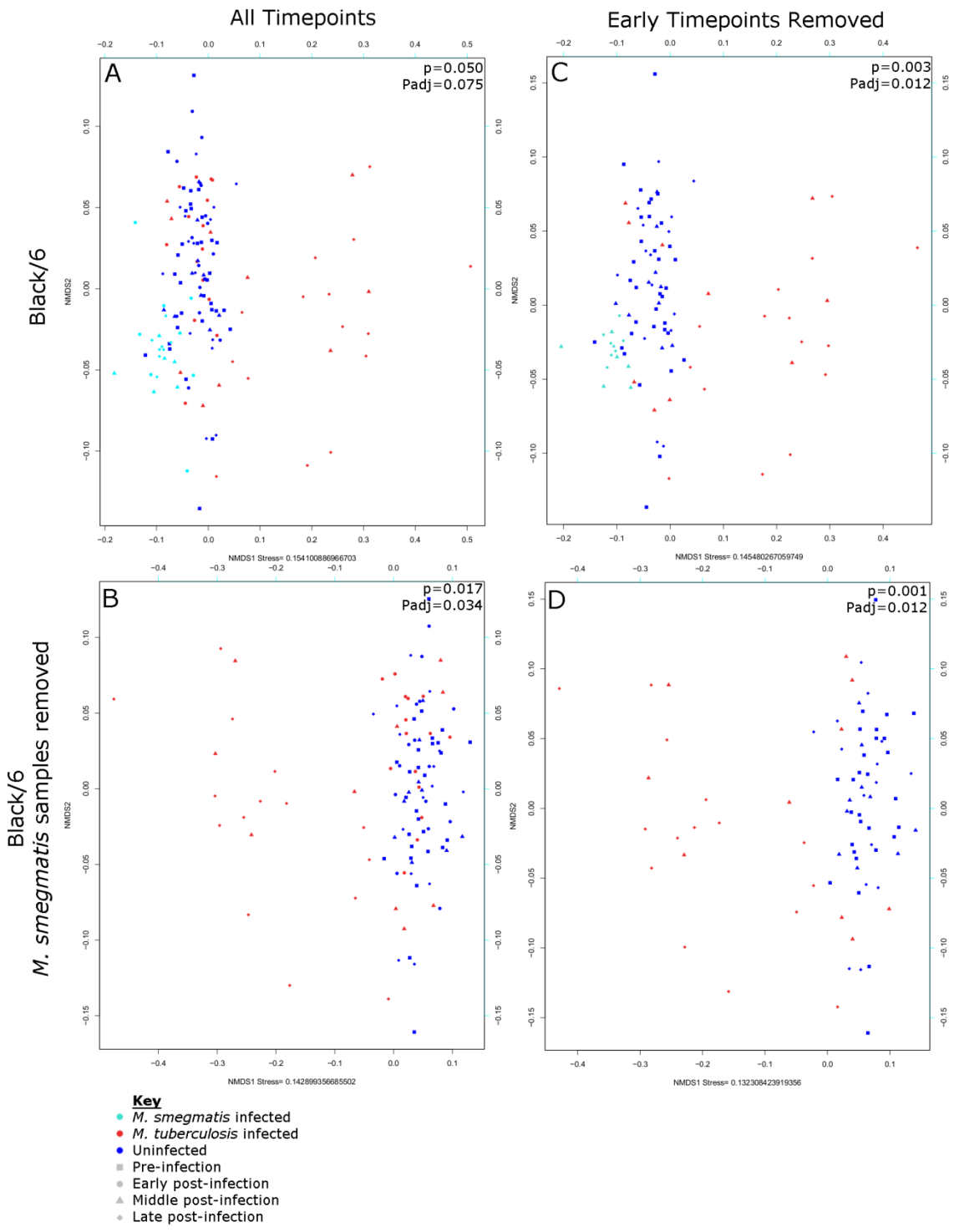


Figure 3-21. Gut gene content of Black/6 mice, as predicted by PICRUSt from 16S sequencing, changes in response to mycobacterial infection.

Black/6 16S sequencing samples colored by infection status and shaped by timepoint. A) All Black/6 samples. B) Black/6 samples with *M. smegmatis* infected mice removed. C) Black/6 samples with early post-infection timepoints removed. D) Black/6 samples with early post-infection samples and *M. smegmatis* infected mice removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

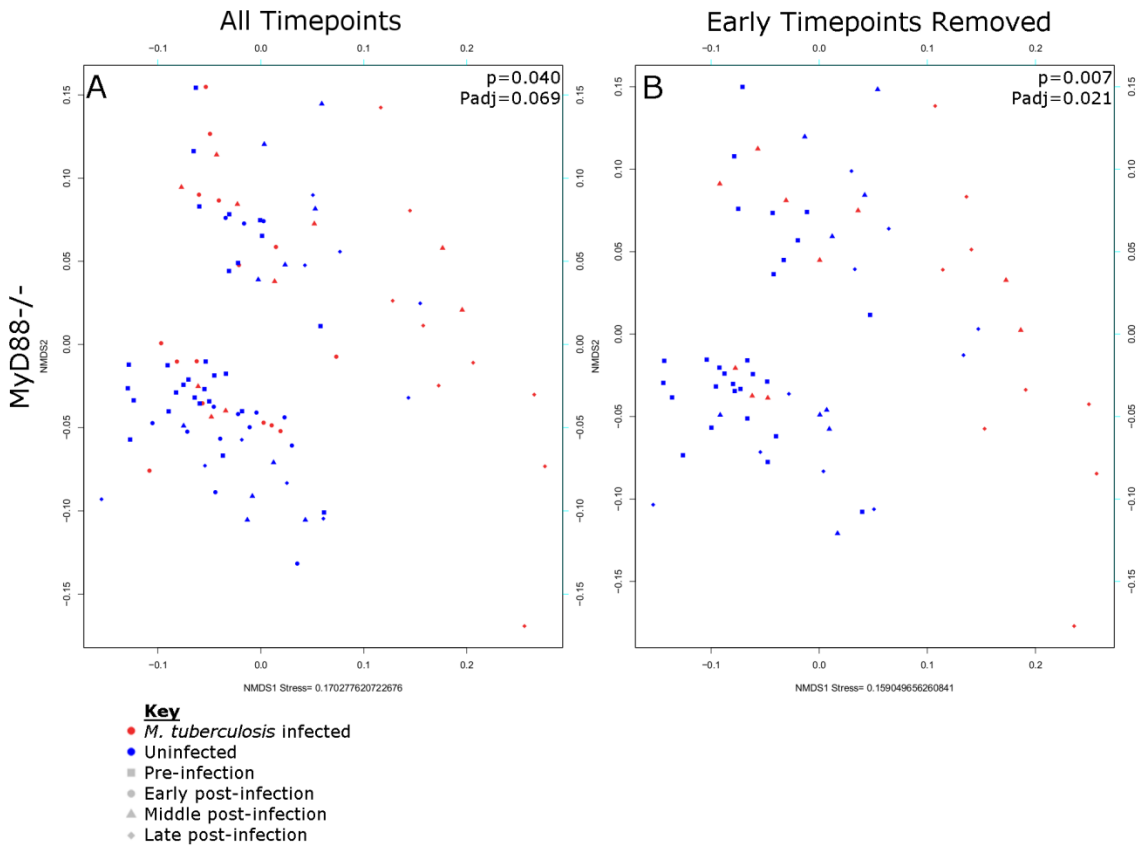


Figure 3-22. Gut gene content of MyD88$^{-/-}$ mice, as predicted by PICRUSt from 16S sequencing, changes in response to *M. tuberculosis* infection by day 10 post-infection.

MyD88$^{-/-}$ 16S sequencing samples colored by infection status and shaped by timepoint. A) All MyD88$^{-/-}$ samples. B) MyD88$^{-/-}$ samples with early post-infection timepoints removed. P values

are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

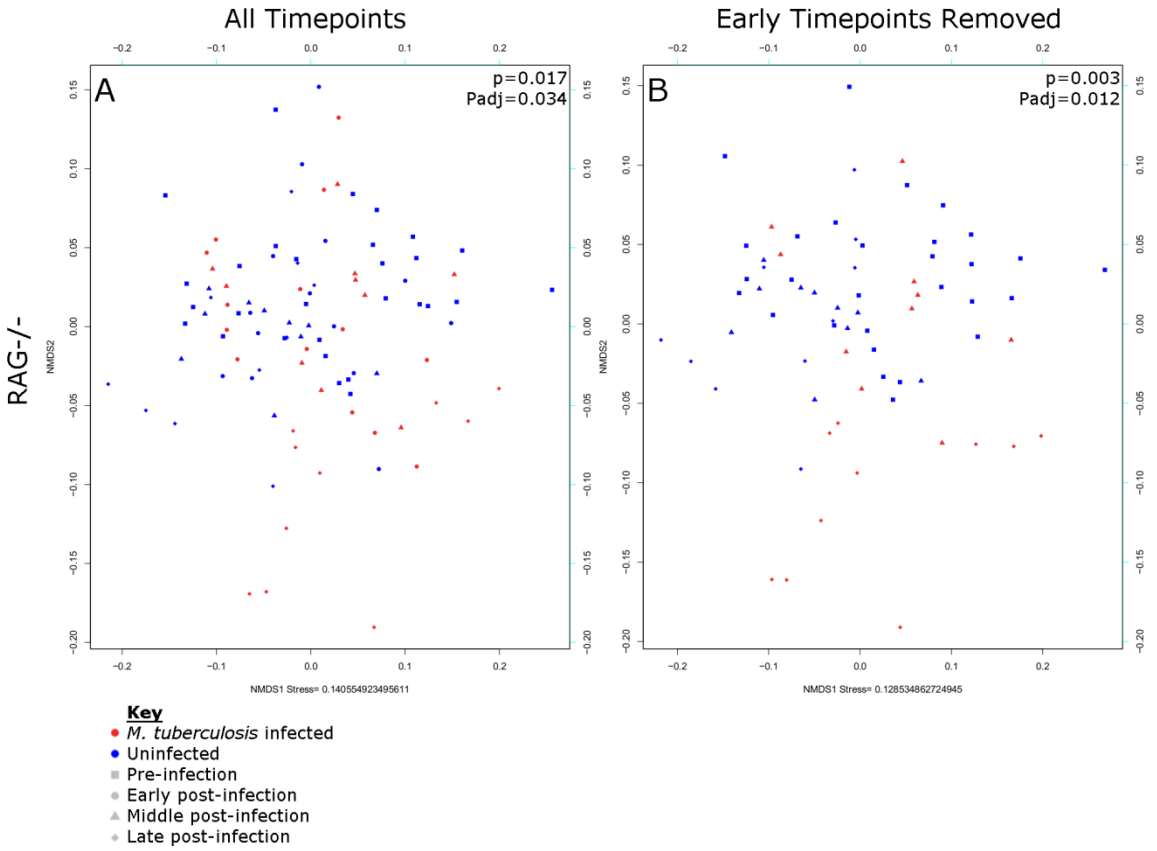


Figure 3-23. Gut gene content of RAG$^{-/-}$ mice, as predicted by PICRUSt from 16S sequencing, changes in response to *M. tuberculosis* infection.

RAG $^{-/-}$ 16S sequencing samples colored by infection status and shaped by timepoint. A) All RAG $^{-/-}$ samples. B) RAG $^{-/-}$ samples with early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

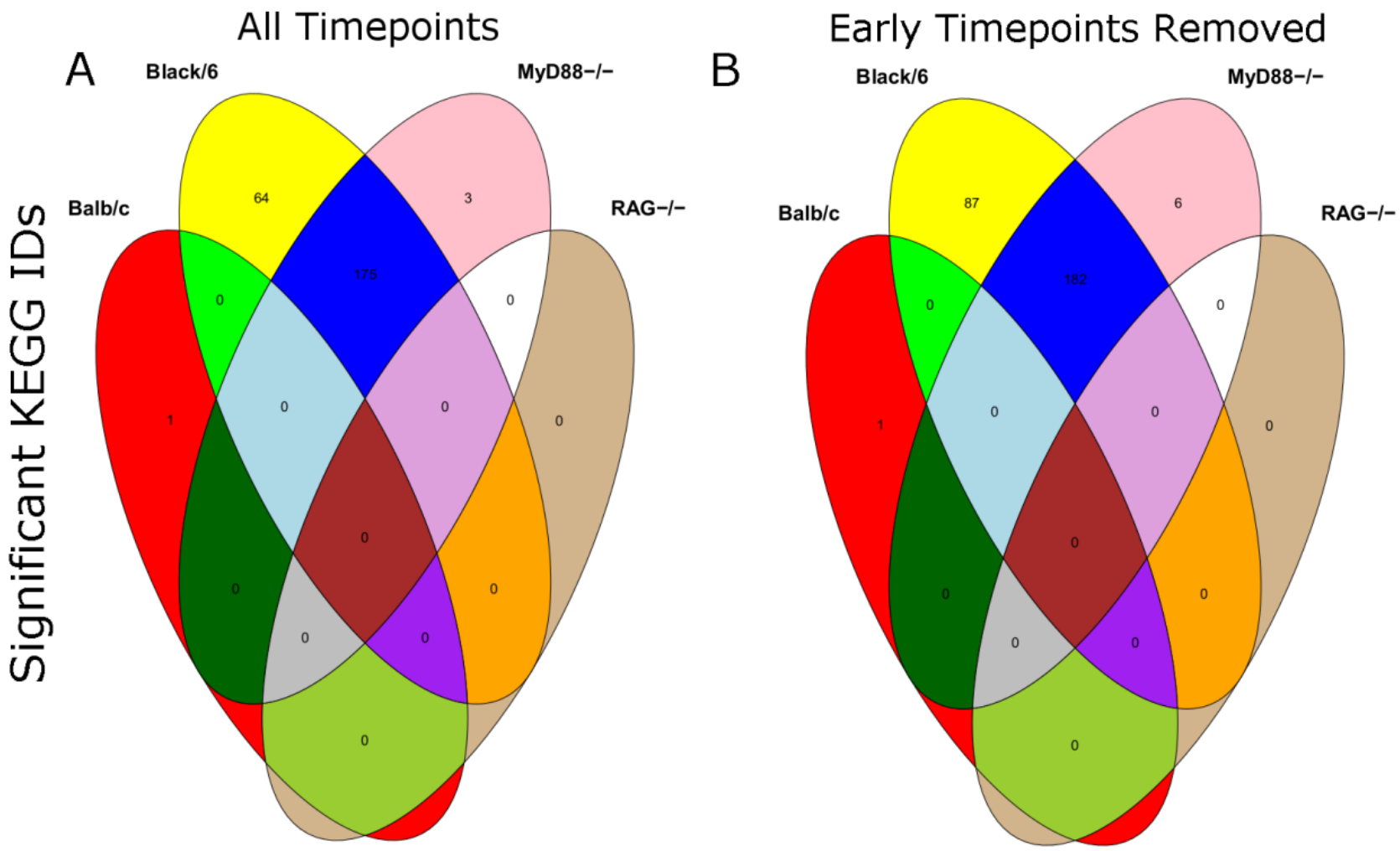

All Timepoints
A
Black/6
MyD88−/−
Balb/c
RAG−/−
64
3
175
0
0
1
0
0
0
0
0
0
0
0
0
0
Early Timepoints Removed
B
Black/6
MyD88−/−
Balb/c
RAG−/−
87
6
182
0
0
1
0
0
0
0
0
0
0
0
0
0
Significant KEGG IDs

Figure 3-24. Overlap between genotypes of KEGG IDs from 16S sequencing significantly different between *M. tuberculosis* infected and uninfected samples.

Venn diagram of overlaps of significant KEGG IDs for each genotype with all timepoints (A) or with early timepoints removed (B).

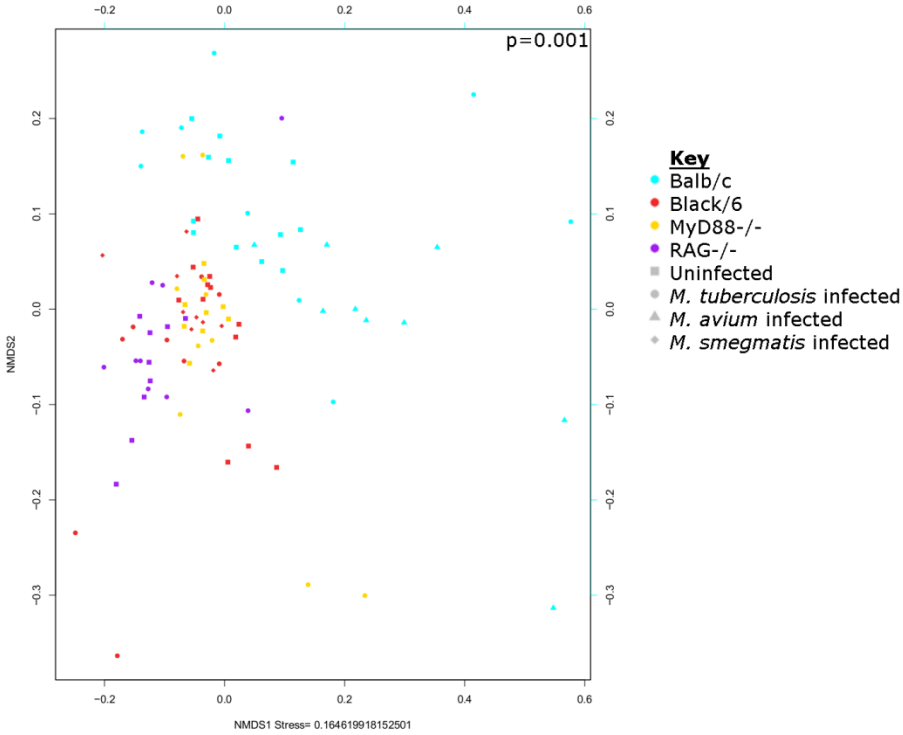


Figure 3-25. Gut gene content as predicted by HUMAnN from whole genome sequencing is significantly different between mouse genotypes.

All samples from whole genome sequencing, colored by genotype and shaped by infection status. P value is for genotype, as calculated by adonis in Vegan and stratified by infecting organism.

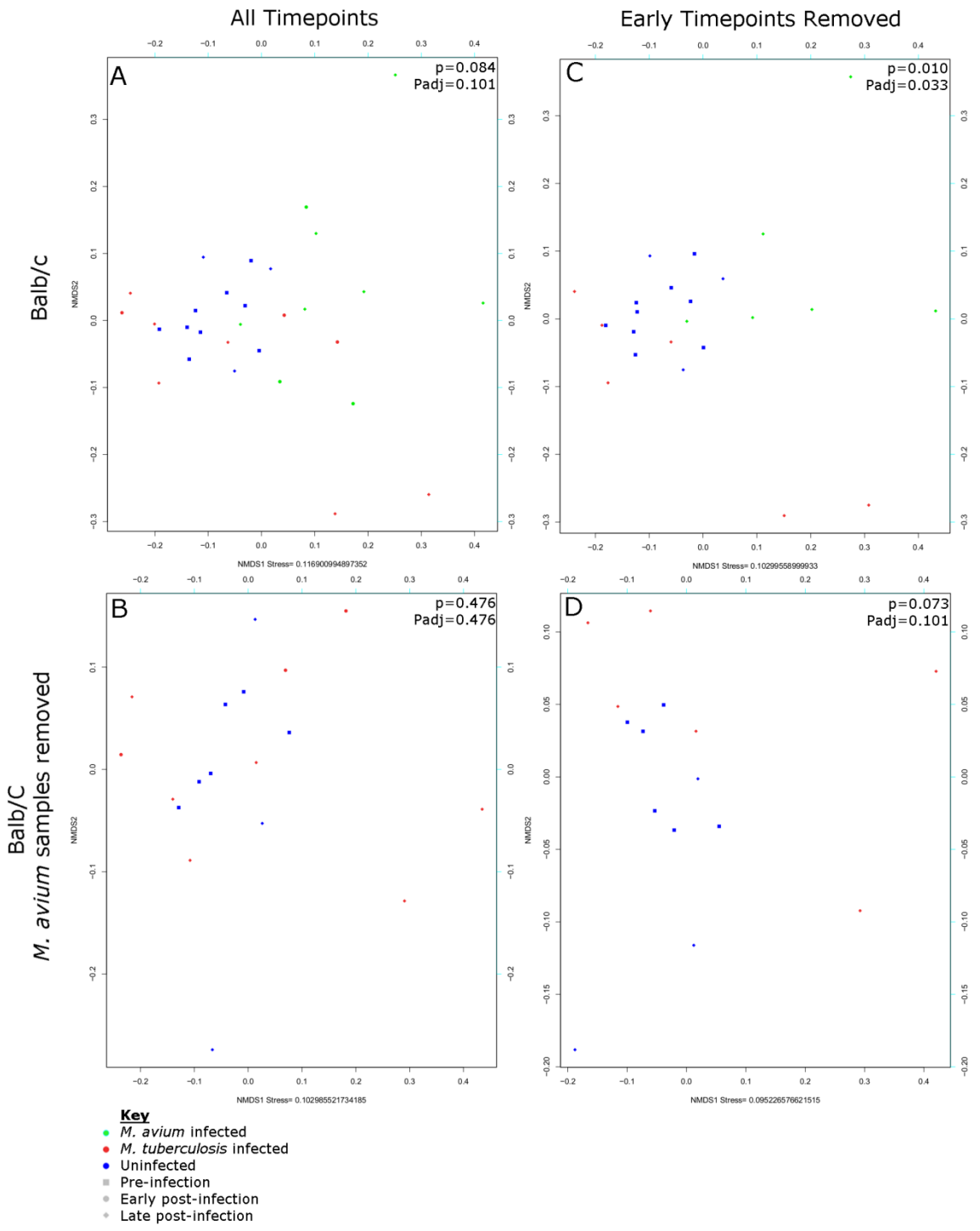


Figure 3-26. Gut gene content of Balb/c mice, as predicted by HUMAnN from whole genome sequencing, does not respond to mycobacterial infection.

Balb/c whole genome sequencing samples colored by infection status and shaped by timepoint. A) All Balb/c samples. B) Balb/c samples with *M. avium* infected mice removed. C) Balb/c samples with early post-infection timepoints removed. D) Balb/c samples with early post-infection samples and *M. avium* infected mice removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

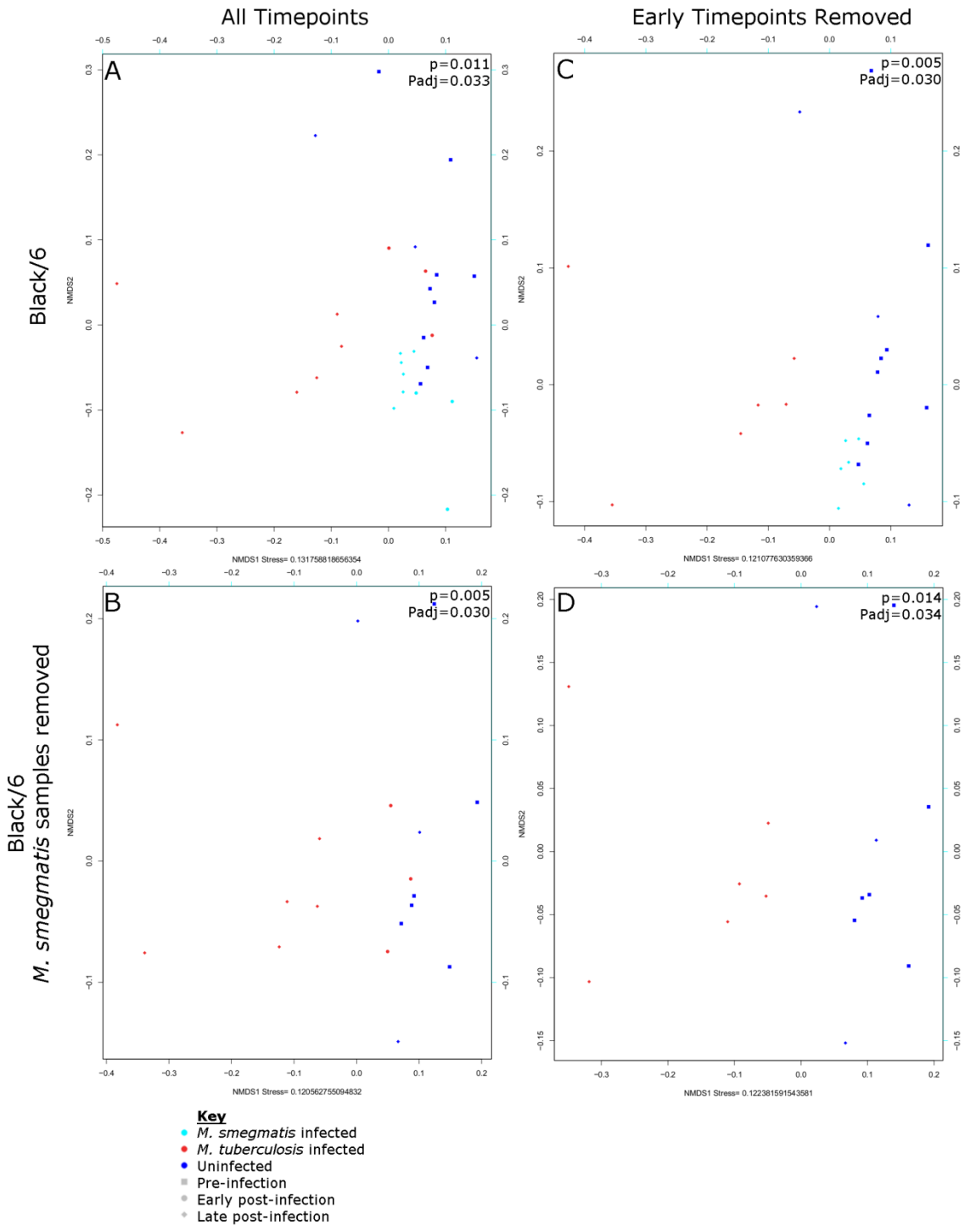


Figure 3-27. Gut gene content of Black/6 mice, as predicted by HUMAnN from whole genome sequencing, responds to mycobacterial infection.

Black/6 whole genome sequencing samples colored by infection status and shaped by timepoint. A) All Black/6 samples. B) Black/6 samples with *M. smegmatis* infected mice removed. C) Black/6 samples with early post-infection timepoints removed. D) Black/6 samples with early post-infection samples and *M. smegmatis* infected mice removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

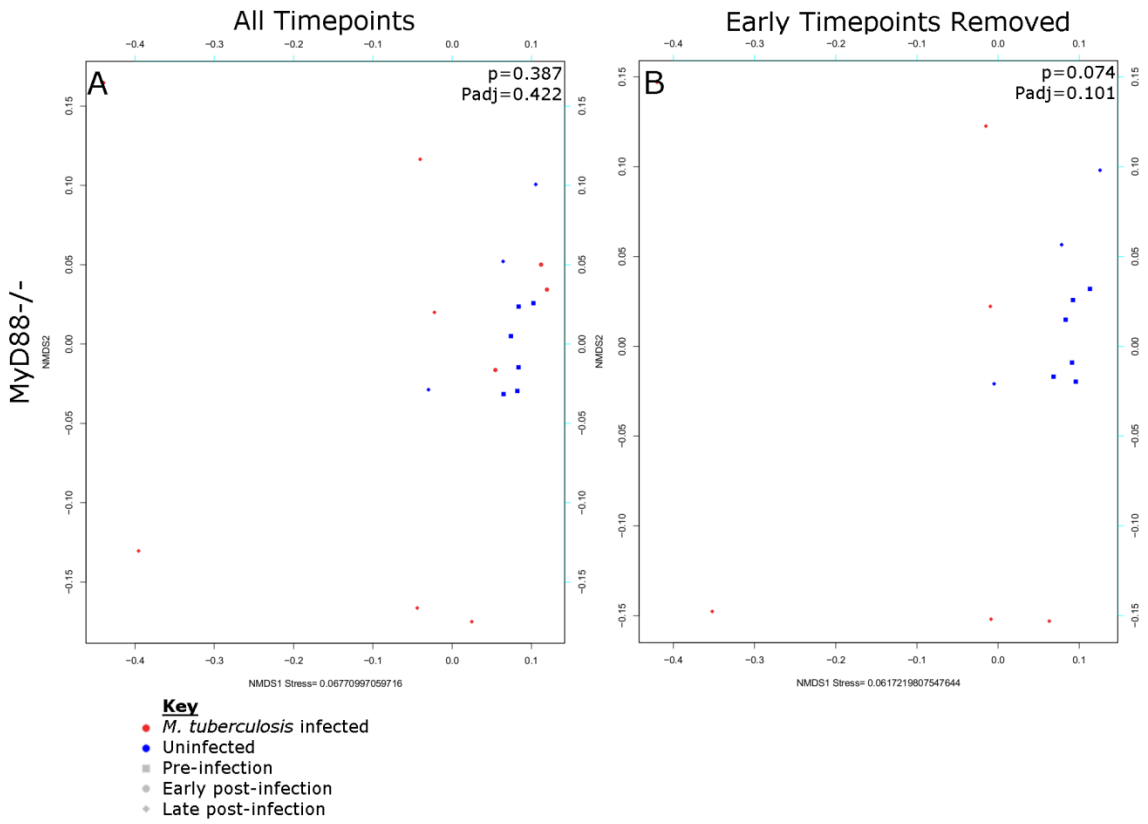


Figure 3-28. Gut gene content of MyD88$^{-/-}$ mice, as predicted by HUMAnN from whole genome sequencing, does not respond to *M. tuberculosis* infection.

MyD88$^{-/-}$ whole genome sequencing samples colored by infection status and shaped by timepoint. A) All MyD88$^{-/-}$ samples. B) MyD88$^{-/-}$ samples with early post-infection timepoints

removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

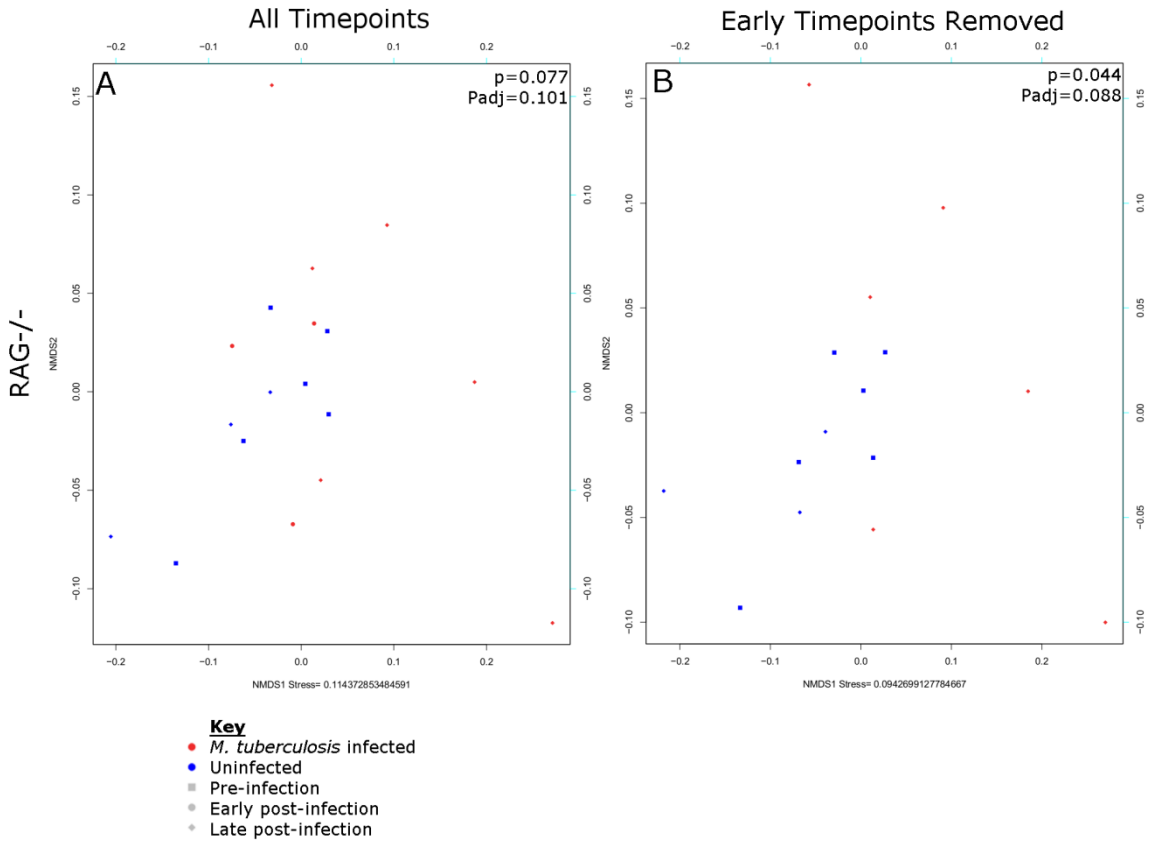


Figure 3-29. Gut gene content of RAG$^{-/-}$ mice, as predicted by HUMAnN from whole genome sequencing, does not respond to *M. tuberculosis* infection.

RAG $^{-/-}$ whole genome sequencing samples colored by infection status and shaped by timepoint. A) All RAG$^{-/-}$ samples. B) RAG$^{-/-}$ samples with early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan, and stratified by cage number. Adjusted P-value was calculated using FDR.

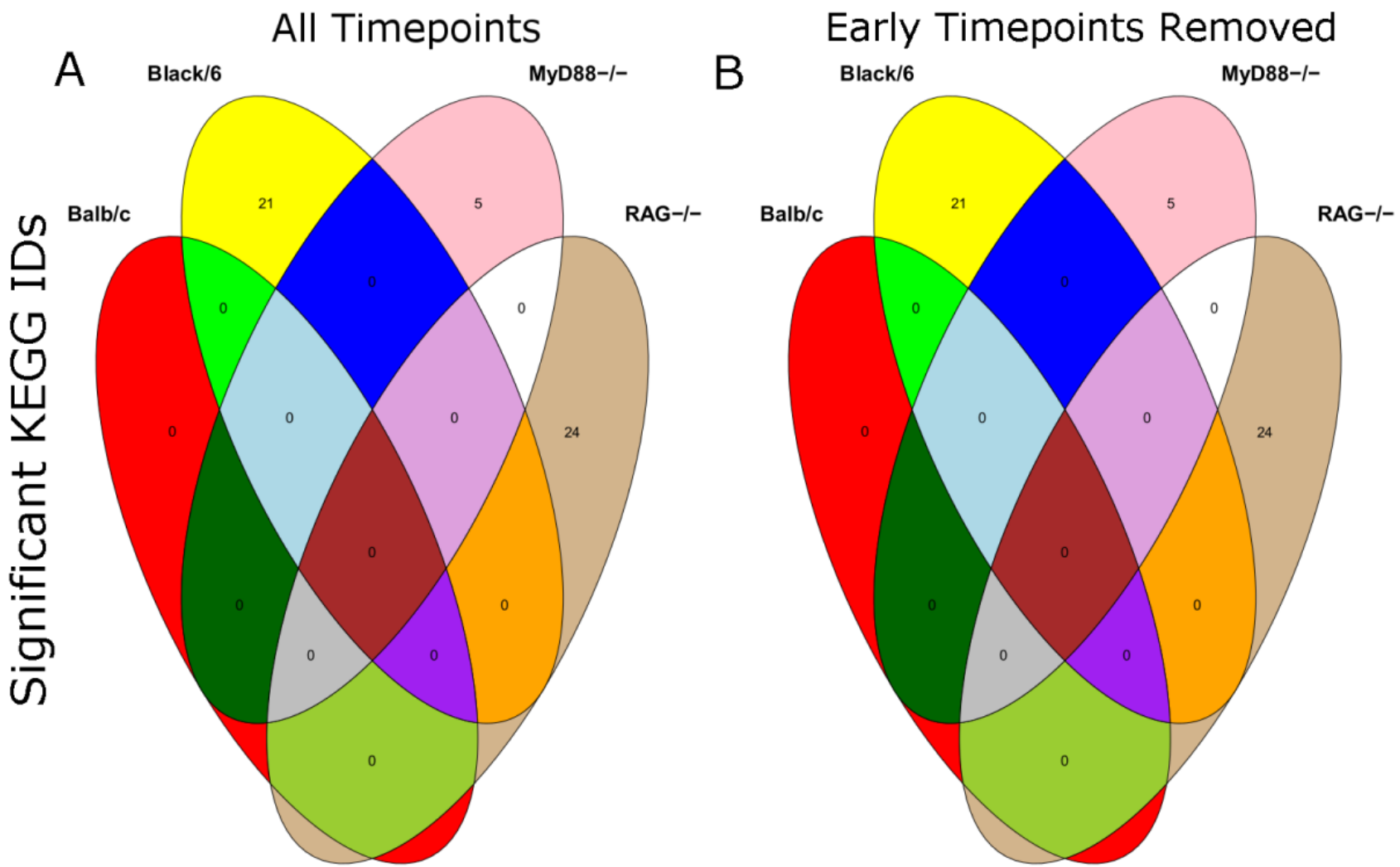

All Timepoints
A
Black/6
MyD88−/−
Balb/c
RAG−/−
21
5
0
0
0
0
0
0
0
24
0
0
0
0
0
0
0
Early Timepoints Removed
B
Black/6
MyD88−/−
Balb/c
RAG−/−
21
5
0
0
0
0
0
0
0
24
0
0
0
0
0
0
0
Significant KEGG IDs

Figure 3-30. Overlap between genotypes of KEGG IDs from whole genome sequencing significantly different between *M. tuberculosis* infected and uninfected samples.

Venn diagram of overlaps of significant KEGG IDs for each genotype with all timepoints (A) or with early timepoints removed (B).

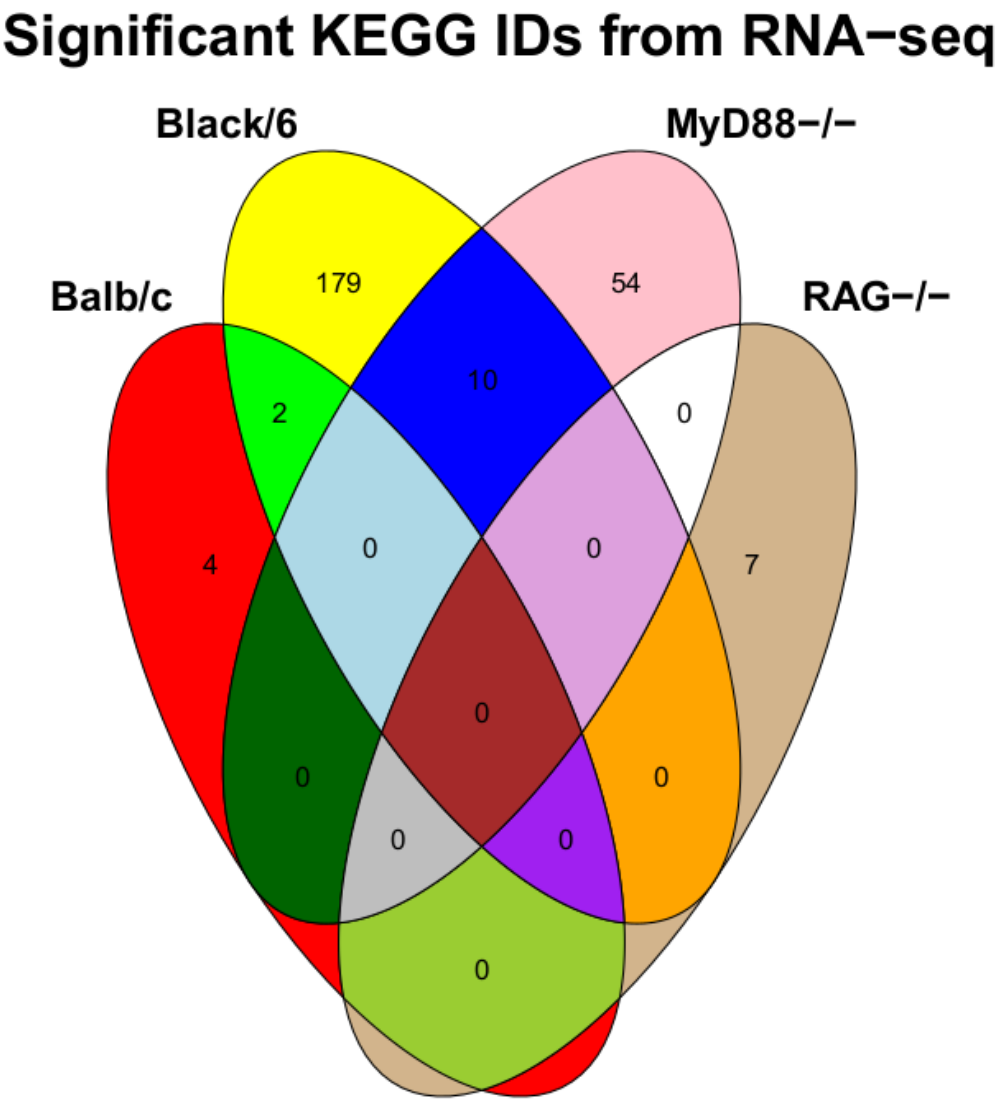


Figure 3-31. Changes in gut microbiota gene expression with mycobacterial infection.

The overlap between KEGG orthologs significantly changed in expression in at least one timepoint compared to day -3.

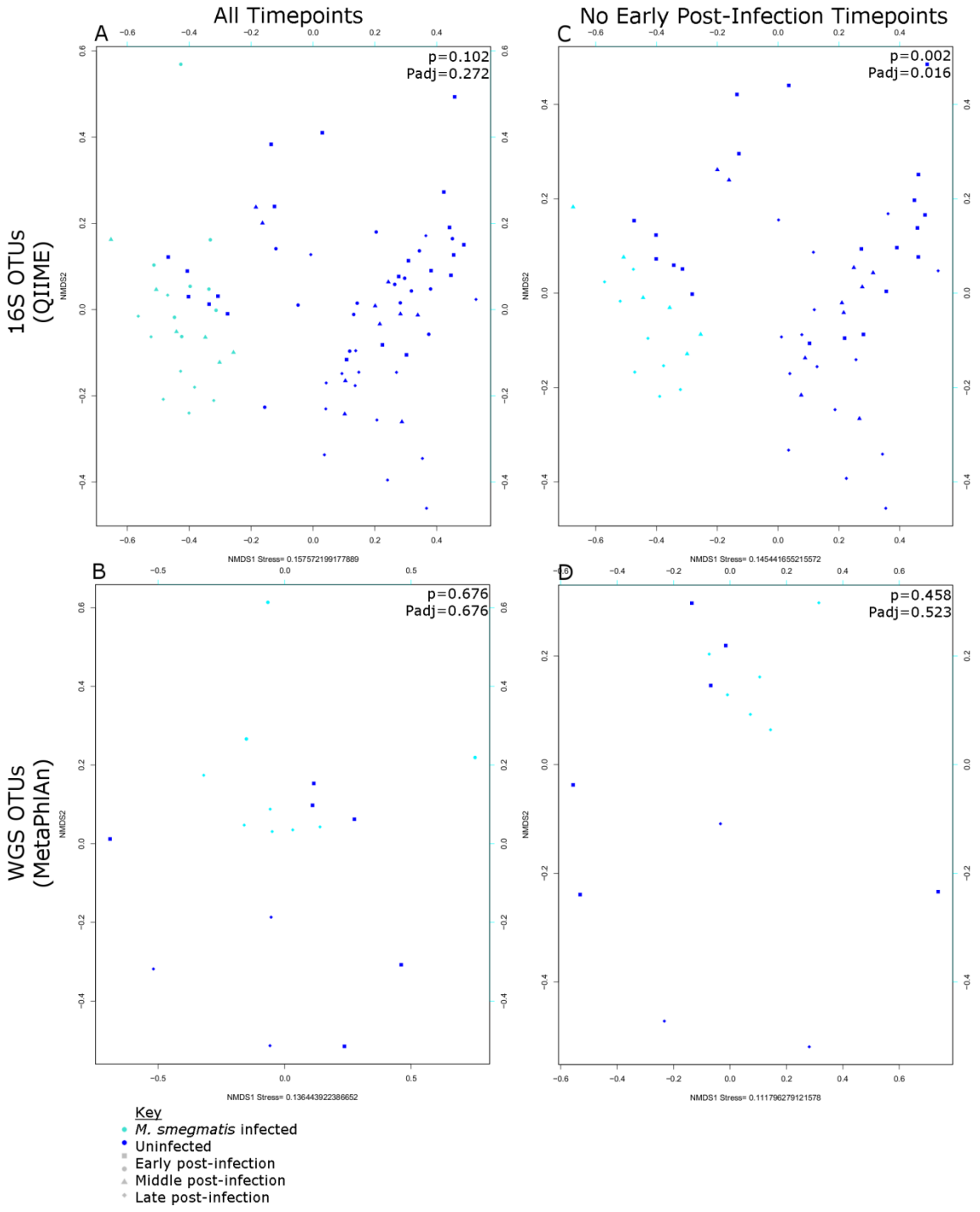


Figure 3-32. Black/6 *M. smegmatis*-infected samples are not different in OTU composition from uninfected samples.

Comparison of uninfected and *M. smegmatis*-infected Black/6 samples. NMDS plots from (A,C) 16S OTU relative abundances from QIIME and (B,D) whole genome sequencing (WGS) OTU relative abundances from MetaPhlAn. A and B include all timepoints while C and D have the early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan. Adjusted P-value was calculated using FDR.

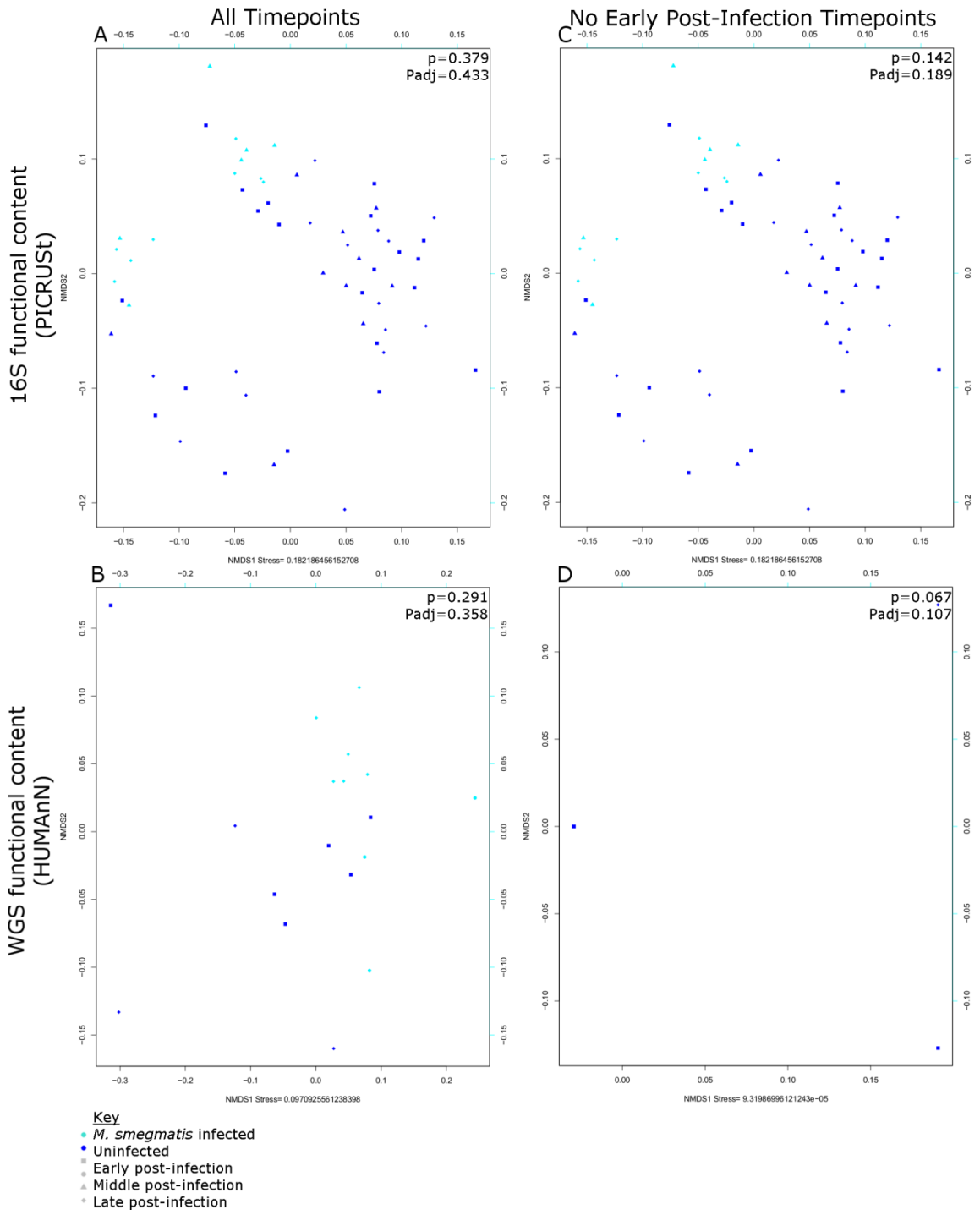


Figure 3-33. Black/6 *M. smegmatis*-infected samples are not different in functional content from uninfected samples.

Comparison of uninfected and *M. smegmatis*-infected Black/6 samples. NMDS plots from (A,C) 16S functional content from PICRUSt and (B,D) whole genome sequencing (WGS) functional content from HUMAnN. A and B include all timepoints while C and D have the early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan. Adjusted P-value was calculated using FDR.

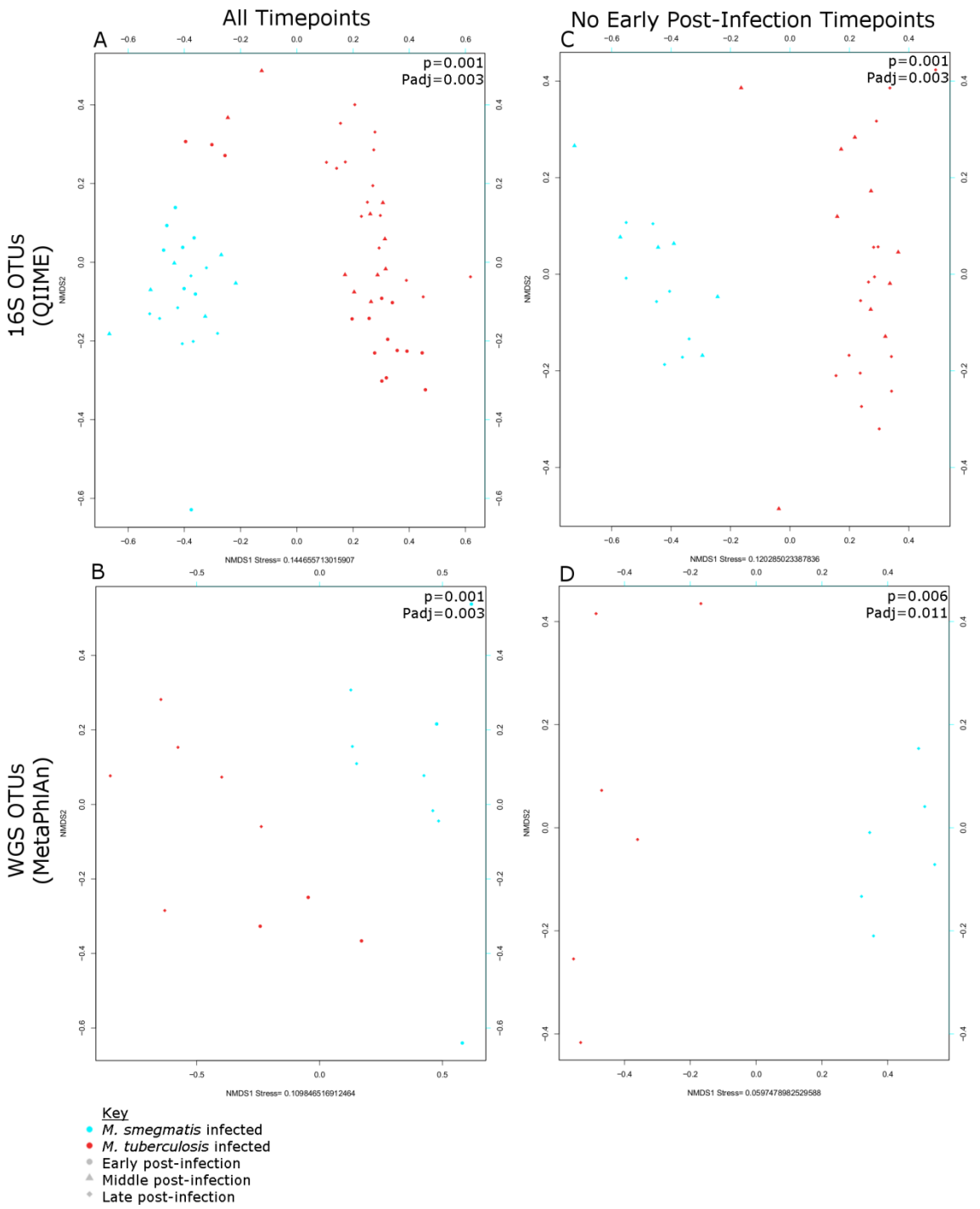


Figure 3-34. Black/6 *M. tuberculosis* infected samples are significantly different from *M. smegmatis* infected samples in OTU composition.

Comparison of *M. tuberculosis* and *M. smegmatis* infected Black/6 samples. NMDS plots from (A,C) 16S OTU relative abundances from QIIME and (B,D) whole genome sequencing (WGS) OTU relative abundances from MetaPhlAn. A and B include all timepoints while C and D have the early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan. Adjusted P-value was calculated using FDR.

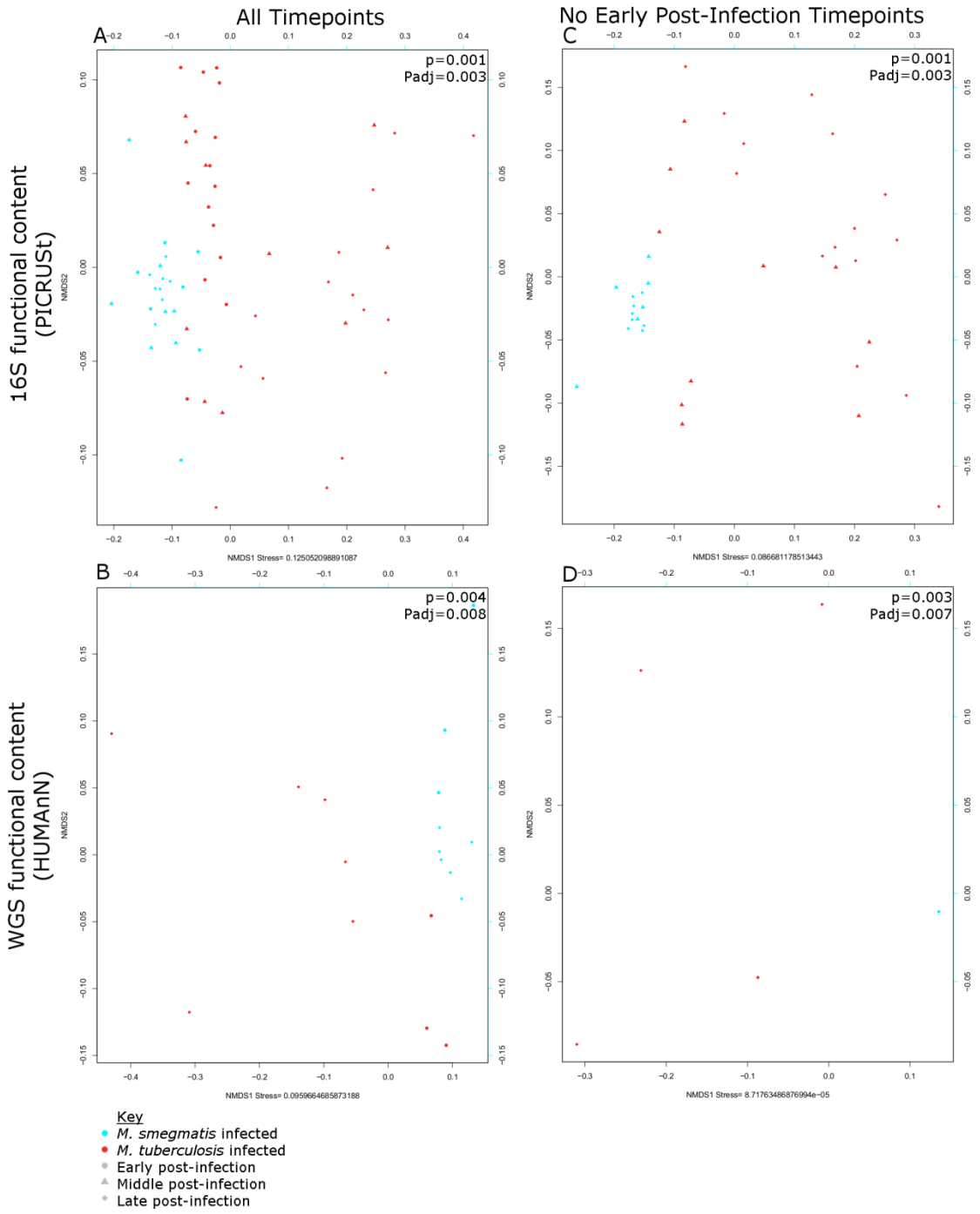


Figure 3-35. Black/6 *M. tuberculosis* infected samples are significantly different from *M. smegmatis* infected samples in functional content.

Comparison of *M. tuberculosis* and *M. smegmatis* infected Black/6 samples. NMDS plots from (A,C) 16S functional content from PICRUSt and (B,D) whole genome sequencing (WGS)

functional content from HUMAnN. A and B include all timepoints while C and D have the early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan. Adjusted P-value was calculated using FDR.

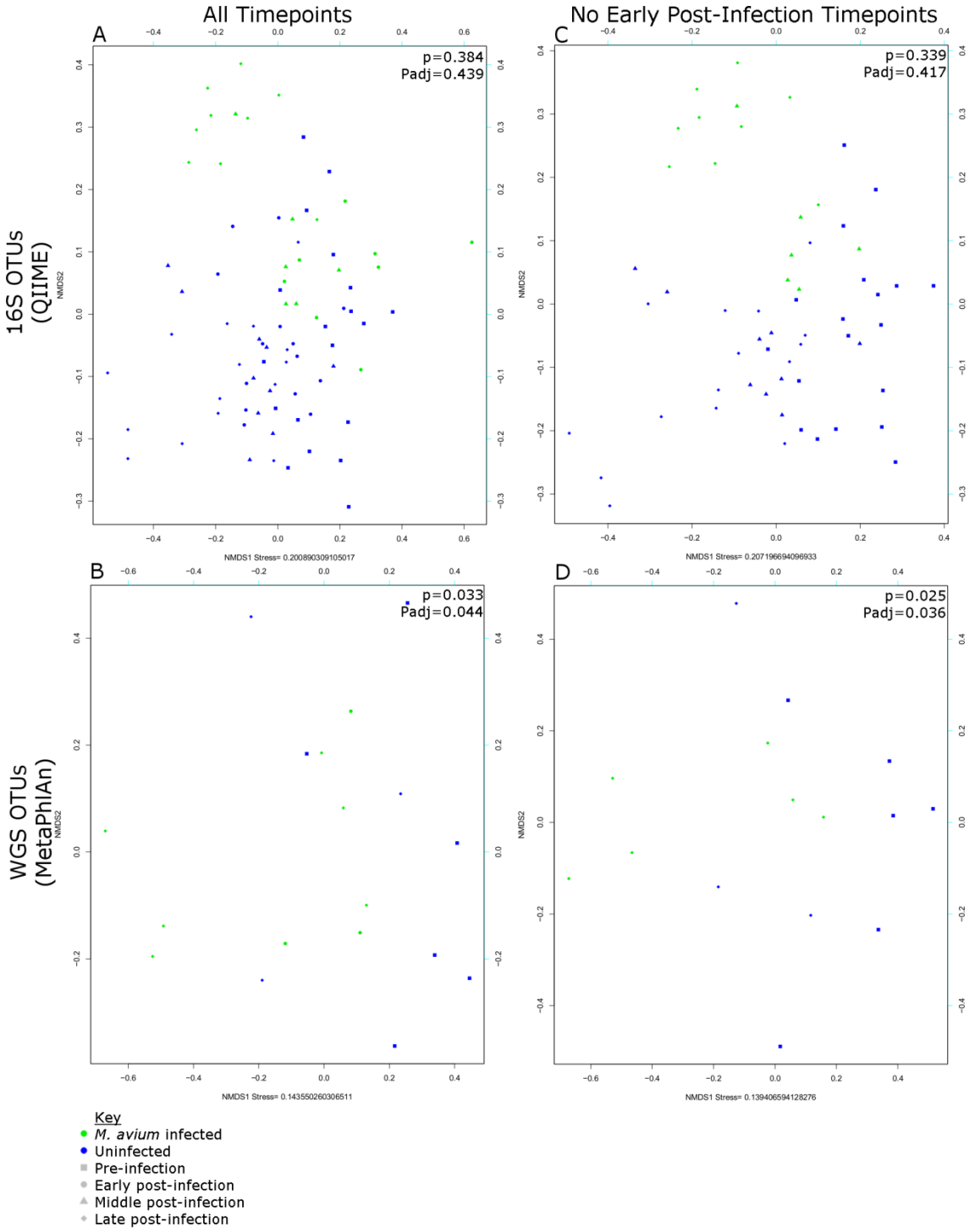


Figure 3-36. Comparison of OTU composition of Balb/c *M. avium* infected samples and uninfected samples.

Comparison of uninfected and *M. avium* infected Balb/c samples. NMDS plots from (A,C) 16S OTU relative abundances from QIIME and (B,D) whole genome sequencing (WGS) OTU relative abundances from MetaPhlAn. A and B include all timepoints while C and D have the early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan. Adjusted P-value was calculated using FDR.

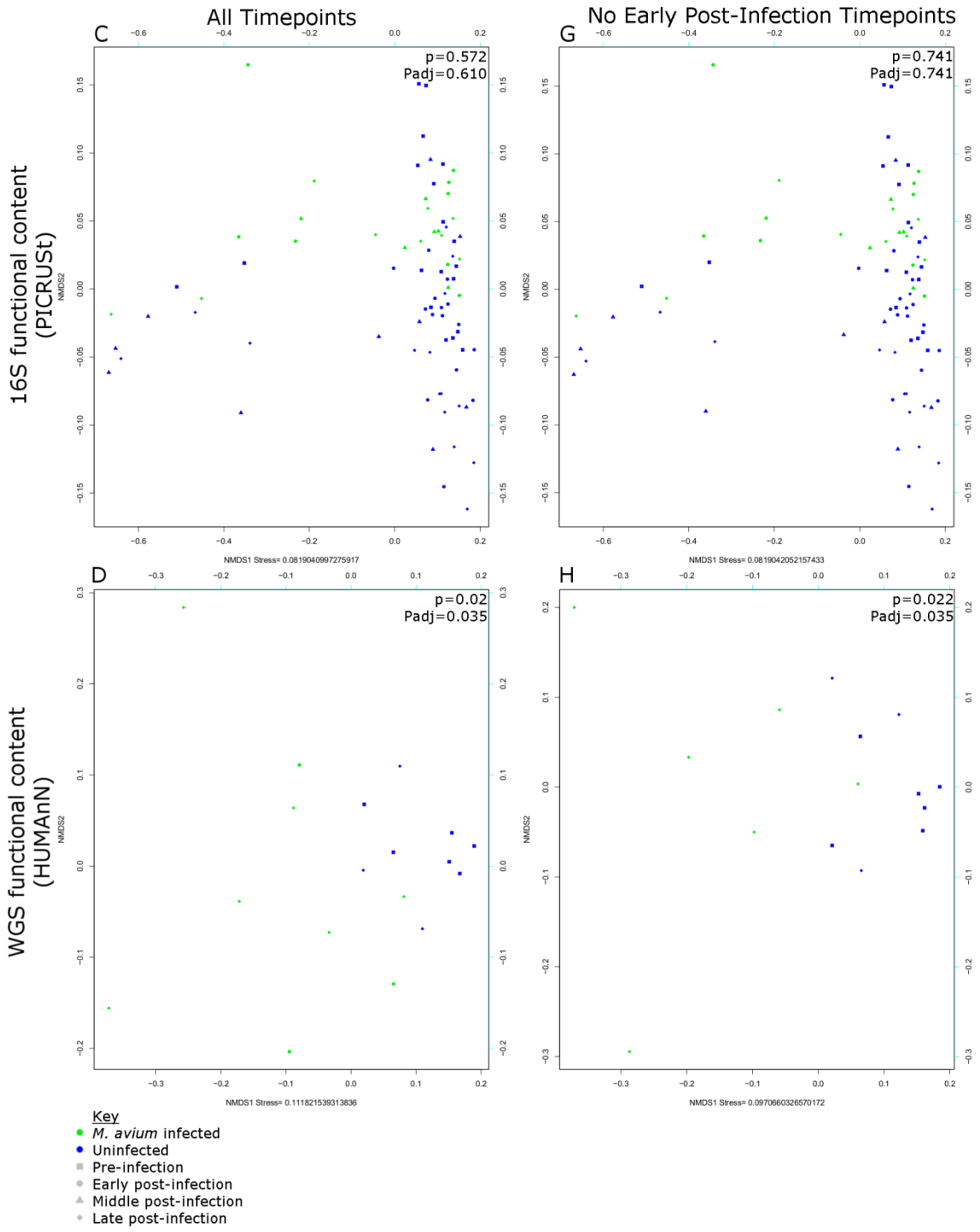


Figure 3-37. Comparison of functional content of Balb/c *M. avium* infected samples and uninfected samples.

Comparison of uninfected and *M. avium* infected Balb/c samples. NMDS plots from (A,C) 16S functional content from PICRUSt and (B,D) whole genome sequencing (WGS) functional content from HUMAnN. A and B include all timepoints while C and D have the early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan. Adjusted P-value was calculated using FDR.

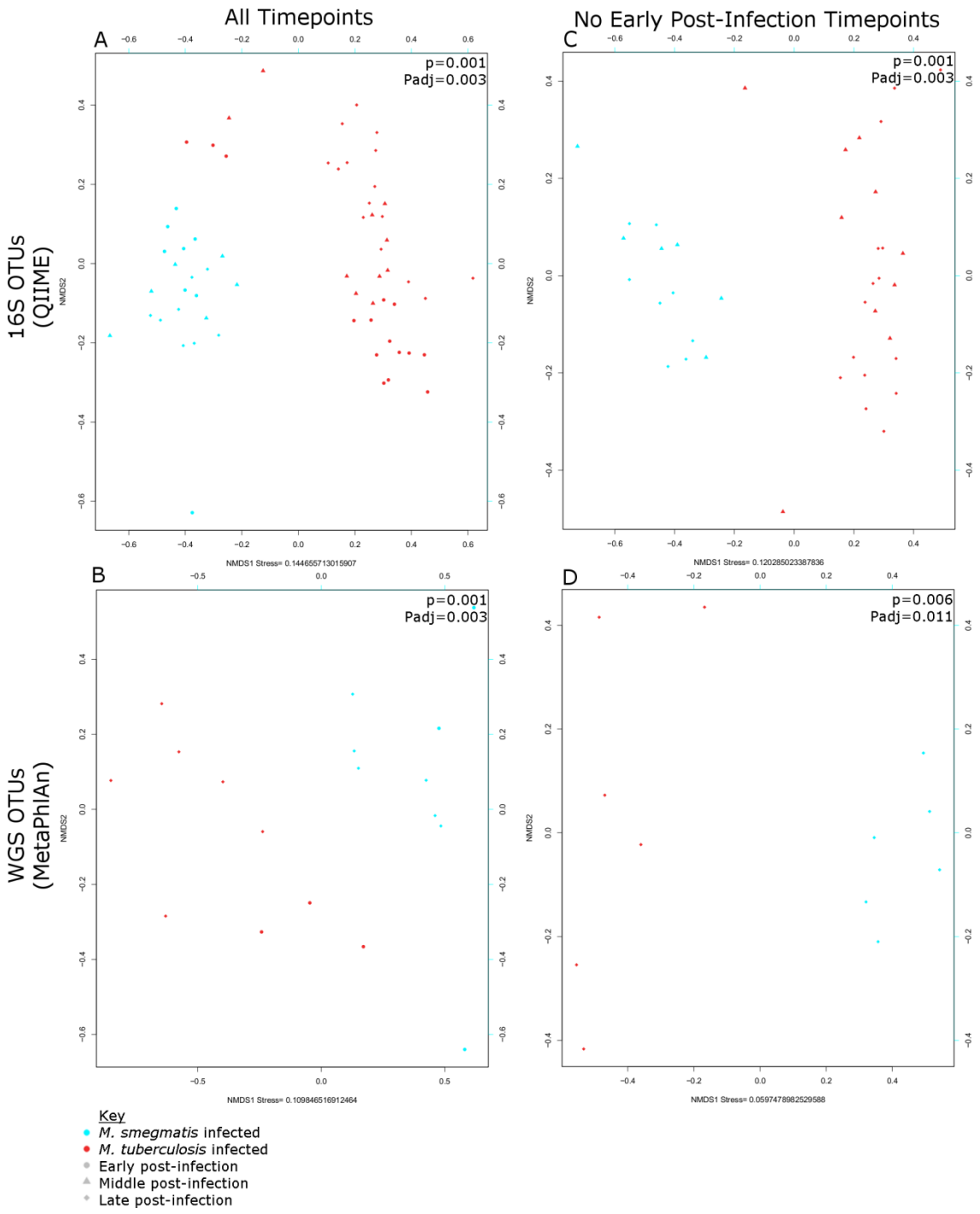


Figure 3-38. Balb/c *M. tuberculosis* infected samples are significantly different from *M. avium* infected samples in OTU composition.

Comparison of *M. tuberculosis* and *M. avium* infected Balb/c samples. NMDS plots from (A,C) 16S OTU relative abundances from QIIME and (B,D) whole genome sequencing (WGS) OTU relative abundances from MetaPhlAn. A and B include all timepoints while C and D have the early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan. Adjusted P-value was calculated using FDR.

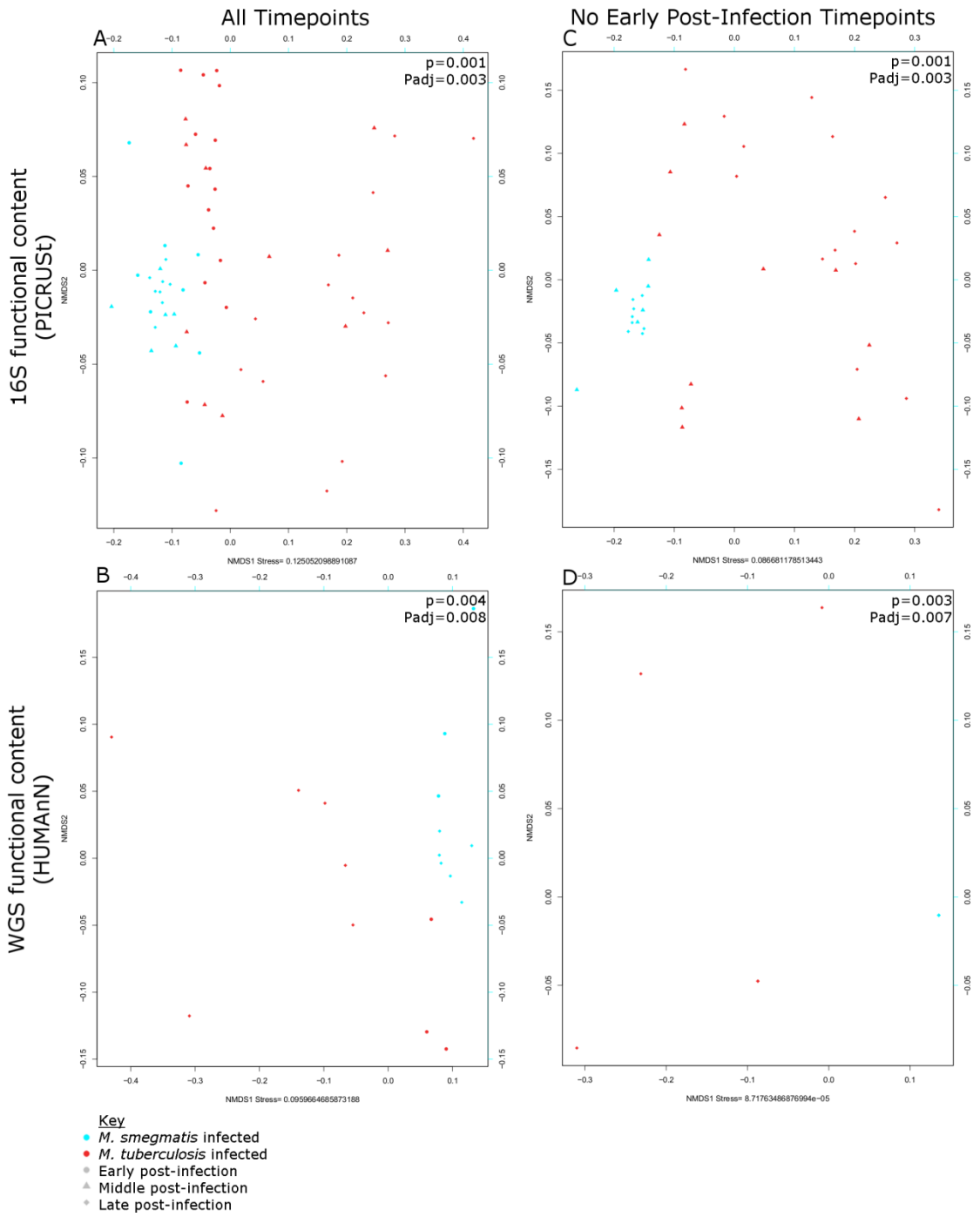


Figure 3-39. Balb/c *M. tuberculosis* infected samples are significantly different from *M. avium* infected samples in functional content.

Comparison of *M. tuberculosis* and *M. avium* infected Balb/c samples. NMDS plots from (A,C) 16S functional content from PICRUSt and (B,D) whole genome sequencing (WGS) functional

content from HUMAnN. A and B include all timepoints while C and D have the early post-infection timepoints removed. P values are for infecting organism as calculated by adonis in Vegan. Adjusted P-value was calculated using FDR.

3.7 Table

Mouse Genotype + Infection Status	Sample	Day -10 (pre)	Day -7 (pre)	Day -3 (pre)	Day 0 (pre)	Day 1 (early)	Day 4 (early)	Day 7 (early)	Day 10 (middle)	Day 14 (middle)	Day 21 (late)	Day 28 (late)	Day 31 (late)
Balb/c *M. tuberculosis* infected	16S sequencing	5	5	5		5	5	5	5	5	5	5	5
	WGS & RNA-seq			3				3			3		3
	CFU					3		3			3		3
	Cytokines					3		3			3		3
Balb/c Uninfected	16S sequencing	5	5	5		5	5	5	5	5	5	5	5
	WGS & RNA-seq			3									3
	Cytokines					3		3			3		3
Black/6 *M. tuberculosis* infected	16S sequencing	5	5	5		5	5	5	5	5	5	5	5
	WGS & RNA-seq			3				3			3		3
	CFU					3		3			3		3
	Cytokines					3		3			3		3
Black/6 Uninfected	16S sequencing	5	5	5		5	5	5	5	5	5	5	5
	WGS & RNA-seq			3									3
	Cytokines					3		3			3		3
RAG-/- M. tuberculosis infected	16S sequencing	5	5	5		5	5	5	5	5	5	5	
	WGS & RNA-seq			3				3			3	3	
	CFU					3		3			3	3	
	Cytokines					3		3			3	3	
RAG-/-	16S	5	5	5		5	5	5	5	5	5	5	

Uninfected	sequencing												
	WGS & RNA-seq			3								3	
	Cytokines					3		3			3	3	
MyD88-/- *M. tuberculosis* infected	16S sequencing	5	5	5		5	5	5	5	5	5	5	
	WGS & RNA-seq			3				3			3	3	
	CFU					3		3			3	3	
	Cytokines					3		3			3	3	
MyD88-/- Uninfected	16S sequencing	5	5	5		5	5	5	5	5	5	5	
	WGS & RNA-seq			3								3	
	Cytokines					3		3			3	3	
Balb/c *M. avium* infected	16S sequencing			3	3	3	3	3	3	3	3	3	3
	WGS & RNA-seq			3				3			3		3
	CFU					3		3			3		3
Black/6 *M. smegmatis* infected	16S sequencing			3	3	3	3	3	3	3	3	3	3
	WGS & RNA-seq			3				3			3		3
	CFU					3		2			2		3

Table 3-1. Samples collected in this experiment.

Numbers in each cell indicate number of mice used for sample collection.

Chapter 4: Mutation of *Rv2887*, a *marR*-like gene, confers *Mycobacterium tuberculosis* resistance to a imidazopyridine-based agent

4.1 Abstract

Drug resistance is a major problem in *Mycobacterium tuberculosis* control, and it is critical to identify novel drug targets and new anti-mycobacterial compounds. We have previously identified an imidazo[1,2-*a*]pyridine-4-carbonitrile-based agent, MP-III-71, with strong activity against *M. tuberculosis*. In this study we evaluated mechanisms of resistance to MP-III-71. We derived three independent *M. tuberculosis* mutants resistant to MP-III-71 and conducted whole genome sequencing of these mutants. Loss of function mutations in *Rv2887* were common to all three MP-III-71-resistant mutants, and we confirmed the role of *Rv2887* as a gene required for MP-III-71 susceptibility using complementation. The *Rv2887* protein was previously un-annotated, but domain and homology analysis suggested it to be a transcriptional regulator in the MarR (multiple antibiotic resistance repressor) family, a group of proteins first identified in *E. coli* to negatively regulate efflux pumps and other mechanisms of multi-drug resistance. We used RNA-seq to identify genes which are differentially expressed in the presence and absence of a functional *Rv2887* protein. One of the genes down-regulated by a functional *Rv2887* protein is *Rv2463,* a putative MarA transcriptional activator. We also found that genes involved in benzoquinone and menaquinone biosynthesis were repressed by functional *Rv2887*. Inactivating mutations of *Rv2887*, a putative MarR-like transcriptional regulator, confer resistance to MP-III-71, an effective anti-mycobacterial compound that shows no cross-resistance to existing anti-tuberculosis drugs.

4.2 Introduction

Tuberculosis (TB) is a devastating disease that infects one third of the world's population and killed 1.5 million people in 2013 [74]. TB is caused by *Mycobacterium tuberculosis* and is challenging and time-consuming to treat. Standard TB treatment is currently six months long and involves a two-month intensive phase consisting of treatment with four antibiotics (isoniazid, rifampin, ethambutol, and pyrazinamide) followed by a four month continuation phase (treatment with isoniazid and rifampin only). However, despite this combination therapy, drug resistance is on the rise, and in 2013, there were an estimated 480,000 cases of multi-drug resistant tuberculosis (MDR-TB), which is defined as bacteria that are resistant to at least isoniazid and rifampin. These cases require up to two years of treatment, but even under these conditions drug resistance is developing, and in 2014, 9% of MDR-TB cases were extensively drug resistant (XDR), meaning that they were also resistant to isoniazid, rifampin, a fluoroquinolone and an injectable anti-TB drug (typically an aminoglycoside) [74]. Thus, there is a great need both for new antibiotics to treat *M. tuberculosis*, and for new drug targets that avoid cross-resistance to currently used therapies.

One such new compound is a imidazo[1,2-*a*]pyridine-4-carbonitrile-based agent, MP-III-71, a compound identified by Pieroni et al. as having an MIC of 0.5 µg/mL against both drug-susceptible and drug-resistant *M. tuberculosis* strains as well as a low Vero cell toxicity of >64 µg/mL (Figure 4-1a) [75]. Thus, this compound is a promising candidate for follow up studies. We were particularly interested in its mechanism, since the susceptibility of MDR and XDR strains to MP-III-71 suggests that it acts against a novel target that can be used against drug resistant strains.

In this study, we generated MP-III-71-resistant *M. tuberculosis* mutants and used whole genome sequencing to identify mutations associated with resistance. We found that all strains had loss of function mutations in *Rv2887*, a small nonessential gene whose only annotation was as a probable transcriptional regulatory protein [76]. Complementation of *Rv2887* restored susceptibility to MP-III-71 and confirmed mutation of *Rv2887* as the mechanism of resistance. Bioinformatics analysis suggested that *Rv2887* belongs to the multiple antibiotic resistance repressor (MarR) family, a family of proteins originally identified in *E. coli* as repressing the expression of a transcriptional activator, MarA, which, when active, increases the expression of efflux pumps and porins, conferring multiple drug resistance [77,78]. Since that time, MarR family proteins have been found in numerous other bacterial species, functioning as both activators and repressors of processes such as drug efflux, virulence factors, catabolic pathways, and response to environmental stresses [79-83]. In addition, we conducted an RNA-seq transcriptome analysis of an MP-III-71 resistant strain to better understand the mechanism of resistance.

4.3 Materials and Methods

4.3.1 MP-III-71

MP-III-71 was synthesized as previously described [75]. It was then dissolved in DMSO at 2048 µg/mL and aliquots were stored in -80 ^{o}C until use. Synthesis of this compound and the N-methyl derivative were performed by Dr. Marco Pieroni.

4.3.2 Synthesis of 2-(4-methoxybenzyl)-3,5-dimethyl-1-oxo-1,5-dihydrobenzo[4,5]imidazo[1,2-*a*]pyridine-4-carbonitrile (the N-Me derivative of MP III-71)

To a suspension of NaH (60% in mineral oil, 22 mg, 0.57 mmol) in dry DMF (2 mL), 2-(4-methoxybenzyl)-3-methyl-1-oxo-1,5-dihydrobenzo[4,5]imidazo[1,2-*a*]pyridine-4-carbonitrile (60 mg, 0.19 mmol), prepared as previously reported,[75] was added at 0 °C. After stirring for 15 min, iodomethane (0.023 mL, 0.38 mmol) was added portion wise and the reaction mixture was refluxed for 4 h. The mixture was poured in ice-water and extracted with ethyl acetate (3 × 10 ml), and the organic layers were washed with water and brine, dried over anhydrous Na_2SO_4 and concentrated under reduced pressure. The crude material was purified through flash chromatography eluting with petroleum ether/ethyl acetate 90/10 to 80/20, yielding the title compound as a white solid. Yield: 76%. Purity: 98 %. ^{1}H-NMR (400 MHz- d_6-DMSO): δ 2.36 (s, 3H), 3.69 (s, 3H), 3.92 (s, 2H), 4.08 (s, 3H), 6.81 (d, *J* = 8 Hz, 2H), 7.15 (d, *J* = 8 Hz, 2H), 7.41 (t, *J* = 8 Hz, 1H), 7.59 (t, *J* = 8 Hz, 1H), 7.75 (d, *J* = 8 Hz, 1H), 8.69 (d, *J* = 8 Hz, 1H); ^{13}C-NMR (100.6 MHz- d_6-DMSO) δ 18.8, 30.8, 31.4, 55.5, 70.0, 110.2, 114.2, 116.1, 116.7, 118.0, 123.1, 127.0, 127.3, 129.3, 132.6, 133.7, 144.8, 147.9, 158.0, 159.2. HRMS (ESI) calculated for $C_{22}H_{19}N_3O_2$ $[M+H]^+$ 358,1477, found: 358,1482.

The title compound was characterized through ^{1}HNMR and ^{13}CNMR. The ^{1}HNMR spectra were recorded on a Bruker 400 Avance spectrometer (400 MHz). ^{13}CNMR spectrum was recorded on a Bruker spectrometer at 100 MHz. Chemical shifts (δ scale) are reported in parts per million (ppm) relative to the central peak of the solvent. ^{1}HNMR Spectra are reported in order: multiplicity and number of protons; signals were characterized as s (singlet), dd (doublet of doublet), t (triplet), m (multiplet), br s (broad signal). HRMS experiments were performed

using a LTQ ORBITRAP XL Thermo by Thermo-scientific instrument coupled to HPLC endowed with a column Alltima C18 5 µ 150mm*4.6mm, Alltech Italia Srl. Reactions were monitored by TLC, on Kieselgel 60 F 254 (DC-Alufolien, Merck). HPLC Method. Column: Alltima C18 5 µ 150mm*4.6mm, Alltech Italia Srl; Flow rate = 1mL/min; isocratic elution for 5 minutes with CH_3CN-H_2O 70% with 0.2% of formic acid; gradient eluation over 25 minutes, from 95% CH_3CN–H_2O with 0.2% of formic acid; isocratic elution for 10 minutes with CH_3CN-H_2O 95% with 0.2% of formic acid.

4.3.3 Mutant isolation

Wild-type *M. tuberculosis* H37Rv was grown in 7H9 broth to stationary phase (OD_{600}=1.98). 500 µL was added to 7H10 plates containing 0.5, 1, 2, 4, or 8 µg/mL (2-32x MIC) of MP-III-71 and incubated at 37 ^{o}C for one month. Colonies grew on all plates, including three colonies on the 8 µg/mL plate. Each of these colonies were grown in 7H9 containing 8 µg/mL of MP-III-71. Of these, one colony did not grow. The other two were streaked onto 7H10 plates with 8 µg/mL MP-III-71 and single colonies were isolated. One colony was picked from each plate, and became mutant 1 and mutant 2. Mutant 3 was selected using the same method but from a biological repeat. In this experiment, the plate with the highest concentration of drug with colonies contained two colonies growing on 1 µg/mL, one large and one small. The large colony became mutant 3.

4.3.4 Drug susceptibility testing

All drug susceptibility testing was performed using the microplate alamar blue assay as previously described [84]. In short, bacteria were added at an OD_{600} of 0.001 to drug dilutions in 7H9 without tween in a 96-well plate. Plates were incubated at 37 ^{o}C for 7 days and then alamar blue was added for 16 hours before plates were read using a fluorescence microplate reader at

544 ex/590 em. Percent inhibition was calculated based on the relative fluorescence units and the MIC was defined as minimum concentration that resulted in at least 90% inhibition. MP-III-71, N-methylated MP-III-71, rifampin (Sigma), and carbonyl cyanide 3-chlorophylhydrazone (CCCP; Sigma) were dissolved in DMSO. Isoniazid (Fluka Analytical), ethambutol (Sigma), kanamycin (Sigma), verapamil hydrochloride (Sigma), chlorpromazine hydrochloride (Sigma), and sodium salicylate (Sigma) were dissolved in water.

4.3.5 DNA extraction

Extraction of genomic DNA was performed on 10 mL cultures in 7H9 broth (mutants were grown in 7H9 with 8 μg/mL MP-III-71) using the CTAB-lysozyme method as previously described [85].

4.3.6 Whole genome sequencing and analysis

For samples run on the Ion Torrent Personal Genome Machine (PGM), 5 μg of genomic DNA were sheared using the Covaris S2 DNA system. The library was then prepared and enriched using the Ion Xpress Plus gDNA Kit and Ion OneTouch Template Kit on the Ion OneTouch, with sequencing performed using the Ion Sequencing Kit v2.0 and 316 chip (Life Technologies). For samples run on the SOLiD, 10 μg of genomic DNA was submitted to the Johns Hopkins SKCC Next Generation Sequencing Center for sequencing. Libraries were constructed according to the protocols provided by Life Technologies (Fragment Library kit) and were run on a High Sensitivity Chip using the Agilent Bioanalyzer to assess size distribution and quality of the amplified library. Quantification of each library was performed by qPCR using the TaqMan Gene Expression Assay as outlined in the Applied Biosystems SOLiD Library Preparation Guide. Libraries were brought to a final concentration of 500 pM and emulsion PCR was performed to

generate 41,000,000 beads to deposit onto an Octet Slide. Sequencing was performed on the Applied Biosystems SOLiD 3 plus System using a single read, 50 bp fragment run.

Reads were aligned to *M. tuberculosis* H37Rv (GI:41353971 from GenBank) using the Burrows-Wheeler Alignment, and SNPs were called using the GATK toolkit [86-88]. Coverage analysis to identify the deletion was performed by comparing the fraction of reads in the parent H37Rv sequence to the fraction of reads in the mutant sequence in each 100bp window of the alignment and identifying the windows with > 40% difference in coverage. From these data, the deletion was picked out as a region with no reads in mutant 1 and an average of 53.6 reads per 100 bp window in the Ion Torrent wild-type H37Rv sequence.

4.3.7 Mutation confirmation

For mutant 1, primers were designed using Primer3 to span the putative deletion [89]. The forward primer was 5'-GTAGCGTGCGAGGTTGAT-3' and the reverse primer was 5'-GAAGCGTTCTTCAGTGGAGT-3', with an expected size of 3753bp in the parent strain and 663bp in the wild-type strain. In addition, primers to span *Rv2887* were used for all three mutants. The forward primer was 5'-AATGCGATGTAGGCTTCAC-3' and the reverse primer was 5'-ATCCACGCCCCAAATATC-3'. Primers were synthesized by Integrated DNA Technologies and diluted to 10 µM in nuclease water. PCR was performed using the 25 µL Taq 2x Master Mix (New England Biolabs), 1 µL forward primer, 1 µL reverse primer, 1 µL genomic DNA, and 22 µL nuclease-free water. The PCR program was 95 ^{o}C for 3 minutes, followed by 30 cycles of 95 ^{o}C for 30 seconds, 53 ^{o}C for 30 seconds, and 1 minute at 68 ^{o}C. After this was 7 minutes at 68 ^{o}C and then products were held at 4^{o}C. PCR products were run on a 1% agarose gel with a 1kb Plus Ladder (Life Technologies), and then the bands were purified using the QIAquick gel extraction kit (QIAGEN). The purified product was diluted to 1.75 ng/µL and submitted with the forward

and reverse primers to Genewiz for Sanger sequences. BLAST was used to align the resulting sequences against Rv2887 to confirm the presence of the mutation [90].

4.3.8 Complementation

Primers were designed using Primer3 to include all of *Rv2887* and 500 bp upstream of the start of *Rv2887* [89]. The forward primer was 5'-GGTATAGGTACCGGTCACGCCTACCACTTG-3', which included a cut site for KpnI, and the reverse primer was 5'-ATCCTCTCTAGATGATGCTCTCGGCTGATAC-3', which included a cut site for XbaI, with an expected size of 1089bp. The primers were dissolved in nuclease-free water and diluted to 10 µM. PCR was performed using 1 µL of each primer, 1500 ng of genomic DNA from *M. tuberculosis* H37Rv, 25 µL Taq 2x Master Mix (New England Biolabs), and nuclease-free water to bring the volume to 50 µL. The resulting mix was thermocycled at 95 ^{o}C for 3 minutes, followed by 30 rounds of 95 ^{o}C for 30 seconds, 58 ^{o}C for 30 seconds, and 60 ^{o}C for 2 minutes. This was finished with 7 minutes at 68 ^{o}C. The resulting product was purified using the QIAquick PCR purification kit (QIAGEN).

Restriction digest was performed on 2 µg of the pMH94h plasmid and 2 µg of the purified PCR product [91]. The digest consisted of 0.2 µL KpnI-HF (New England Biolabs), 0.2 µL XbaI (New England Biolabs), 5 µL NEB buffer 4, and 44.6 µL DNA and nuclease-free water. The mix was incubated at 37 ^{o}C for 4 hours. The digested DNA was run on 1% agarose gel and the bands extracted using the QIAquick gel extraction kit (QIAGEN). The digested plasmid was then treated with alkaline phosphatase, calf intestinal (CIP) (New England Biolabs). 50 ng of CIP-treated digested plasmid was ligated with 150 ng of digested PCR product using the Quick Ligation Kit from New England Biolabs. The ligated product was used to transform *E. coli* One Shot TOP10 electrocompetent cells (Life Technologies), and the transformed bacteria were

plated on LB agar containing 150 µg/mL Hygromycin B (Roche). Plates were incubated overnight at 37 ^{o}C and four single colonies were picked the next day and grown in 5 mL LB broth with 150 µg/mL Hygromycin B. The next day, plasmid was isolated using the QIAprep spin miniprep kit (QIAGEN).

To confirm plasmid sequence, 600 ng of purified plasmid were digested as described above, and the digest run on 1% agarose to confirm presence of the insert. All four products were submitted to Genewiz for Sanger sequencing, along with the forward primer 5'-AGCGCATAGGAACGATTTAC-3' and the reverse primer 5'- ACCCGGTAGAGCAGATAGC-3'. BLAST was used to confirm the correct sequence in all four plasmids; one was chosen randomly to continue [90].

Each strain was grown to around OD_{600} 0.8 in 50 mL of 7H9 and made electrocompetent. In brief, the culture was spun down for 10 minutes at 37 ^{o}C at 4500 RPM and was re-suspended in 30 mL of 10% glycerol (Sigma). This was repeated 4 times, halving the volume used for re-suspension each time, until the bacterial were finally re-suspended in 2.5 mL 10% glycerol. 100 µL of cells were incubated at 55 ^{o}C for 5 minutes and then the bacteria were electroporated with 25 ng/µL plasmid in 50 µL water using 330 µF / 375 Volts / 8 KOhms. The cells were then transferred into 2 mL of 7H9 and incubated at 37 ^{o}C for 48 hours. The transformed cells were then centrifuged at 10,000 RPM for 5 minutes, re-suspended in 200 µL and the entire volume plated on 7H11 with 50 µg/mL Hygromycin B. Plates were incubated 37 ^{o}C for four weeks.

4.3.9 Transposon mutants

Transposon mutants were available through the TARGET project (http://webhost.nts.jhu.edu/target/).

4.3.10 RNA extraction

Total RNA was extracted from 50 mL of culture at around 1 OD_{600}. *M. tuberculosis* cultures were centrifuged and the bacterial pellet was resuspended in TRIzol (Invitrogen). This mixture was transferred to 1.8 mL O-ring tubes containing 0.5 mL of 0.1 mm glass beads (BioSpec Products). Cells were incubated at 25 ^{o}C for 10 minutes, lysed by six cycles of bead-beating for 30 seconds and cooling on ice for 1 minute, using a mini-beadbeater at 4,800 RPM. Lysed cells were centrifuged for 5 minutes at 13,000 RPM, the supernatant was transferred to a fresh microfuge tube and RNA was then extracted as described [92,93]. The quality of RNA was assessed using a Nanodrop (ND-1000, Labtech) and Agilent 2100 Bioanalyzer (Agilent Technologies).

4.3.11 Quantitative reverse-transcription PCR

15 µg of RNA was treated with DNase I (New England Biolabs) according to manufacturer's instructions. Reverse transcription was performed using the iScript Reverse Transcriptase (BioRad) on 1 µL of DNase-treated RNA. The primers used for *Rv2887* for qRT-PCR were 5'-GTTCGCTACCGGCTACATTG-3' (forward) and 5'-CTAGTCGGACCCGAGCTTCT-3' (reverse). The primers for our control housekeeping gene, *sigA*, were 5'-CCATCCCGAAAAGGAAGACC-3' (forward) and 5'- TCGAGGTCTGGTTCAGCGTC-3' (reverse). *sigA* is a housekeeping gene whose expression remains constant and is commonly used for *M. tuberculosis* [94]. qRT-PCR was performed on 2 µL cDNA using the iQ SYBR Green Supermix (BioRad) on the Applied Biosystems StepOnePlus using 95 ^{o}C for 3 minutes, followed by 40 cycles of 95 ^{o}C for 15 seconds and 53 ^{o}C for 1 minute, then a melt curve from 55-95 ^{o}C. The average fold change compared to wild-type, normalized to *sigA*, were calculated using data from two technical replicates of three independent experiments.

4.3.12 Bioinformatics analysis

BLAST alignment was performed using the online tool [90]. Consensus sequences for marR and marA were downloaded from the conserved domain database [95]. Secondary structures were predicted using the default settings for the online versions of Phyre2, PROMALS3D, and PRALINE [96,97]. All tools made the same prediction, and the results from Phyre2 are shown in Figure 4-4. Several alignment tools were used, including PRALINE, MUSCLE, and ClustalW, all on the default settings [98-100]. The alignment from PRALINE is the one depicted in Figure 4-4. The MarR protein sequences to align were selected from CDD, using the top five most diverse members, including the consensus sequence [95]. The GI numbers for these sequences were 192988597 (*Salmonella typhimurium* SlyA), 75341253 (*Enterococcus hirae* NapB), 81637589 (*Bacillus subtilis* YvnA), and 81703996 (*Bacillus subtilis* YhjH). The online tool for SIFT was used to assess the effect of non-synonymous mutations [101].

4.3.13 RNA-seq and analysis

The strains were grown to an OD_{600} of 0.6 in 150 mL, and MP-III-71 or an equivalent volume of DMSO was added at to a final concentration of 0.0625 μg/mL MP-III-71. Cultures were incubated shaking at 37 ^{o}C for 6 hours and then RNA extraction was performed as described above. Ribosomal RNA was removed using the Ambion MICROBExpress Bacterial mRNA Enrichment Kit (Life Tehcnologies) and the results checked using the Agilent 2100 Bioanalyzer (Agilent Technologies). Library preparation was performed using the Ion Total RNA-Seq Kit v2 (Life Technologies). Samples were barcoded using the Ion Xpress RNA barcodes and each replicate was pooled into one library. Template preparation was performed with the Ion PGM Template OT2 200 Kit using the Ion OneTouch 2 and Ion OneTouch ES (Life Technologies). Sequencing was performed on the Ion Personal Genome Machine using the Ion 318 chip with

the Ion PGM Sequencing 300 Kit with weighted buckets (Life Technologies). Reads were aligned to *M. tuberculosis* H37Rv (GI:41353971 from GenBank) using the Burrows-Wheeler Alignment [87]. Differential expression was assessed using DESeq, EdsgeR and Cufflinks, all of which gave the same results [102-104]. Results from DE-Seq are presented here.

4.3.14 Nucleotide accession numbers

All sequencing data has been deposited in the GenBank Sequence Read Archive under BioProject PRJNA280011. Individual accession numbers are given in Table 4-1.

4.4 Results

4.3.1 Whole genome sequencing reveals that *Rv2887* mutations confer MP-III-71 resistance

Spontaneous mutants resistant to MP-III-71 were selected by growing *M. tuberculosis* H37Rv on 7H9 agar plates containing up to 8 μg/mL of the compound, a concentration which is 32x the MIC (see Materials and Methods). A total of three mutant colonies were identified during two separate selections, and their resistance to MP-III-71 was confirmed using microplate alamar blue assay (MABA), as shown in Table 4-2. Each mutant had an MIC of 1-2 μg/mL, which is 2-8x the MIC of the parent strain, which showed an MIC of 0.25-0.5 μg/mL.

Genomic DNA from each of the three mutants and the wild-type parent H37Rv strain was submitted for deep sequencing. Mutant 1 and 2 and wild-type H37Rv were sequenced on an Ion Torrent Personal Genome Machine (PGM), with an average of 1.4 million reads and an average read length of 111 base-pairs (bp; Table 4-1). The parent H37Rv strain was re-sequenced using an ABI SOLiD instrument, along with mutant 3, with an average of 25.2 million 50 bp reads (Table 4-1). Reads were aligned to the *M. tuberculosis* H37Rv reference genome,

and mutations identified in either of the two parental H37Rv sequences were removed from further analysis. After this analysis, mutant 1 had 7 SNPs and 11 indels, mutant 2 had 9 SNPs and 4 indels, and mutant 3 had 1 SNP (Table 4-3). Using the TARGET collection of mutants, we tested the MP-III-71 susceptibility of *M. tuberculosis* CDC1551 strains with a transposon inserted in *Rv2658c* or *Rv3668c*, the only two genes mutated in our study that had an available transposon mutant [105]. However, neither of these mutants had an altered susceptibility to MP-III-71 (Table 4-4).

Interestingly, *Rv2887* was mutated in both mutant 2 and 3, with a 2 bp deletion in mutant 2 and a non-synonymous mutation in mutant 3, both of which were confirmed with Sanger sequencing of PCR amplified DNA fragments (Table 4-2). In addition, we analyzed the coverage of each strain, and found a 3,183 bp deletion in mutant 1, which includes *Rv2887* (Table 4-2, Figure 4-2a). The deletion was confirmed by PCR with primers inside and outside the deleted area (Figure 4-2b), and the outside PCR product was submitted for Sanger sequencing to identify the exact boundaries of the deletion. From this, we determined that the deletion was from position 3,194,362 to position 3,197,545 in the H37Rv reference nucleotide sequence, resulting in partial deletion of *Rv2885c* and *Rv2888c,* and full deletion of *Rv2886c* and *Rv2887*. Thus, we hypothesized that loss of function of *Rv2887* confers resistance to MP-III-71.

4.3.2 Complementation of *Rv2887* confirms that mutation of this gene confers resistance to MP-III-71

Wild-type *Rv2887* was cloned into an integrating pMH94-derived plasmid, along with 500 bp upstream to capture its native promoter [91]. The plasmid was introduced into each of the three mutants and MABA was performed with MP-III-71. In addition, in mutant 1, which had a full deletion of *Rv2887*, real-time PCR was performed to confirm that expression of *Rv2887*

from the plasmid is the same as wild-type levels (Figure 4-3). Introduction of the wild-type version of *Rv2887* was sufficient to restore susceptibility to MP-III-71 to wild-type levels in all three mutants (Table 4-2), confirming that loss of function of *Rv2887* results in resistance to MP-III-71.

4.3.3 Comparative analysis of *Rv2887* with known MarR transcriptional regulators

BLAST analysis revealed that *Rv2887* belongs to the MarR Pfam family, Pfam01047, with an E-value of 6.38×10^{-13} [90,106,107]. MarR homologs contain a winged helix-turn-helix motif, and they have been shown to play a role either as transcriptional repressors or activators of several different pathways, including response to antibiotic and oxidative stress [108,109]. The *E. coli mar* (multiple antibiotic resistance) locus, which has been well characterized, consists of two transcriptional units, *marC* and *marRAB*, which are under the control of a centrally located promoter region between these two divergent operons. MarR, which has been shown to function as a repressor in *E. coli,* binds to repeats within the operator and prevents transcription until bound by certain chemical compounds, such as tetracycline, chloramphenicol, and sodium salicylate [110]. Binding induces a conformational change in MarR, which reduces its affinity for the repressor DNA sequence, allowing transcription of *marC* and *marRAB.* MarA, which functions as a transcriptional activator in *E. coli,* is then expressed, leading to elevated transcription of a diverse regulon of genes [110]. In *E. coli*, this regulon governs numerous functions including up-regulation of efflux pumps and is associated with the multiple antibiotic resistance phenotype.

Alignment of *Rv2887* and the mutant strains against the consensus Pfam01047 sequence and other members of this family showed that valine 61, mutated to alanine in

mutant 3, is a conserved amino acid in MarR-like proteins in multiple species, and thus this small amino acid change may have a significant impact on protein function (Figure 4-4) [95]. Several alignment tools were used, including PRALINE, MUSCLE, and ClustalW to confirm this finding; Figure 4-4 shows the results from PRALINE [98-100]. The importance of this amino acid was confirmed by SIFT, which predicted that the V61A mutation will affect protein function, based on the conservation of amino acid residues [101]. In addition, the 2 bp deletion in mutant 2 changes the four C-terminal residues into 29 residues, increasing the length of the protein, and adding an additional disordered region including two new alpha helices (Figure 4-4c). Thus, the changes in both mutant 2 and mutant 3 are predicted to affect *Rv2887* function.

4.3.4 Efflux inhibitors and resistance to MP-III-71

Based on the role of MarR in regulating efflux pumps in *E. coli*, we tested the efflux pump inhibitors verapamil, chlorpromazine, and carbonyl cyanide 3-chlorophenyl hydrazone (CCCP) [111-114]. However, these compounds had no effect on the MIC of MP-III-71 to wild-type H37Rv, mutant 1 or complemented mutant 1 (Table 4-5). Furthermore, there was no difference in the susceptibility of mutant 1 compared to wild-type for these three compounds. Thus, while *Rv2887* may play a role in drug efflux in *M. tuberculosis*, it appears that it does not govern efflux pumps targeted by these three inhibitors. *M. tuberculosis* H37Rv has an estimated 148 efflux pumps, and so efflux of MP-III-71 may still be controlled by loss of *Rv2887* function, despite the lack of activity of verapamil, chlorpromazine and CCCP [115].

Given the role of MarR in drug resistance in *E. coli* and other organisms, we also tested whether any of the mutants had altered susceptibility to the TB antibiotics isoniazid, rifampin, ethambutol and kanamycin. None of the mutants showed MIC changes to these drugs compared to wild-type (Table 4-5; data not shown for mutants 2 and 3). This correlates with our previous

findings that MP-III-71 is effective against drug resistant clinical isolates, and inhibits a unique pathway not targeted by existing anti-TB drugs.

4.3.5 Transcriptional profiling of an MP-III-71 resistant mutant

Since *Rv2887* is a putative transcriptional regulator, we performed RNA-seq on the Ion Torrent PGM to determine whether it governs the transcription of other genes. We chose to focus on mutant 1 and complemented mutant 1 because mutant 1 had the full deletion of *Rv2887*, and because it had the highest MP-III-71 MIC. We harvested the RNA after 6 hours incubation with 0.0625 µg/mL MP-III-71 (1/4 wild-type MIC) or with an equivalent volume of DMSO only. Triplicate samples were run on the Ion Torrent PGM, with an average of 1.2 million reads and 115 bp read length (Table 4-1).

After performing differential expression analysis comparing the mutant and its complemented strain, we identified six genes that were significantly differentially expressed between mutant and wild-type, including *Rv2887* (Table 4-6). All genes identified as significant were up-regulated in the MP-III-71 resistant mutant compared to the complemented strain, with the exception of *Rv2887* and *Rv2886c. Rv2886c* down-regulation in the mutant is probably due to artifactual overexpression of this gene in the comparator strain, since the end of this gene is at the start of *Rv2887*, and the complementation plasmid contained 500 bp upstream of *Rv2887*. The fact that the other genes are up-regulated with the deletion of *Rv2887* is consistent with the hypothesis that *Rv2887* serves as a transcriptional repressor.

4.3.6 The regulon of *Rv2887* includes genes involved in menaquinone biosynthesis and a potential MarA protein

Three of the significant genes identified in our RNA-seq analysis as being up-regulated upon *Rv2887* deletion were *Rv0560c*, *Rv0559c*, and *Rv0558.* These three genes form a cluster in the H37Rv genome, although *Rv0558* is in the opposite orientation to the other two genes, and many of the reads aligned to this gene were in the other orientation, suggesting read-through from *Rv0559c*. *Rv0559c* encodes a non-essential exported protein found in culture filtrate, membrane and whole cell lysate, and, along with *Rv0560c*, is up-regulated in rifampin-resistant strains and induced by salicylate [76,116-119]. *Rv0560c* is a benzoquinone methyltransferase involved in the biosynthesis of isoprenoid compounds, and it is also up-regulated in iron-limited and anaerobic conditions [76,120-122]. *Rv0558* (*menH*) encodes an S-adenosylmethionine-dependent methyltransferase found in the membrane that catalyzes the final step in menaquinone biosynthesis [76,123,124]. Menaquinone (vitamin K) is an essential electron carrier in the respiratory chain and is particularly important in *M. tuberculosis* survival under low oxygen conditions [124]. The other gene up-regulated by *Rv2887* deletions was *Rv2463*. This gene has been annotated as *lipP*, a probable esterase/lipase [76]. LipP has a low level of long-chain triacylglycerol hydrolase activity, and is induced after 6 hours of nutrient starvation [125].

Given that *Rv2887* is in the MarR family but *M. tuberculosis* lacks the adjacent *marRAB* locus identified in *E. coli*, we used BLAST to seek *E. coli* MarA (accession number EDV65186.1) homologs which may be situated elsewhere in the H37Rv genome [90]. Only 6 genes were identified, one of which was *Rv2463*, with an Expect (E) value of 0.23 (Table 4-7). Although a function for *Rv2463* as a MarA-like transcriptional activator seems unlikely for a protein with lipase activity, especially given the relatively high E value, we used this same method to look for

other MarR-like proteins in H37Rv, with *E. coli* MarR (accession number AAK21292.1) as the query, and picked up *Rv2887*, in addition to 7 other potential MarR-like proteins (Table 4-7). Thus, *M. tuberculosis* H37Rv possesses several putative MarR- and MarA-like proteins. However, further studies will be needed to assess function.

4.3.7 *Rv2887*-dependent susceptibility to MP-III-71 may involve drug methylation

We hypothesized that mutation of *Rv2887* may be indirectly causing resistance to MP-III-71 by altering expression of its true target. We had access to transposon mutants of *Rv0559c* and *Rv2463* [105]. Given that loss of *Rv2887* caused these genes to be up-regulated, we hypothesized that deletion of these genes might result in MP-III-71 hyper-susceptibility. We tested the MIC of MP-III-71 of the *Rv0559c* and *Rv2463* mutants. However, neither of these mutants had an altered susceptibility to MP-III-71, suggesting that their up-regulation with loss of *Rv2887* is not the cause of MP-III-71 resistance (Table 4-4). However, the transposon in *Rv2463* was inserted 2 amino acids from the C terminus, so this mutant may retain some degree of intact function (Table 4-4).

This left altered expression of *Rv0558* and *Rv0560c* as potential reasons for MP-III-71 resistance. *Rv0560c* is up-regulated in the presence of salicylate, which is of particular interest since salicylate is one of the compounds that interferes with the repressor activity of MarR in *E. coli* [126-128]. As a result, we tested the susceptibility of mutant 1 to salicylate, but found that deletion of *Rv2887* had no effect on the MIC of salicylate (all strains had an MIC of 250-500 μg/mL). Sub-MIC levels of salicylate did reduce susceptibility of the mutant, but not the complemented or wild-type strains, to MP-III-71 (Table 4-5).

Since *Rv0558* and *Rv0560c* are methylases, we next hypothesized that MP-III-71 may be methylated, leading to inactivation of the compound. Up-regulation of these genes by altering *Rv2887* function would increase the amount of methylated, inactivated drug. We tested our strains for susceptibility to N-methylated MP-III-71 (Figure 4-1b). Interestingly, this compound had no effect on the growth of any strain, even at the highest concentrations tested (32 μg/mL; Table 4-5). This showed that methylation results in inactivation of MP-III-71 and suggests that one possible mechanism of resistance to this compound may be through methylation of the drug.

4.4 Discussion

The development of new drug targets and new drugs to combat *M. tuberculosis* is critical as rates of drug resistance increase. Here, we explore the function of one potential new anti-mycobacterial compound, MP-III-71, and we show that loss of function mutations in *Rv2887* confer resistance to this compound.

Rv2887 is a transcriptional regulator in the MarR family. Very little is known about the role of the *mar* operon in mycobacteria. McDermott et al. showed that expression of *E. coli* MarA in *M. smegmatis* increases resistance to several antibiotics, including rifampin, isoniazid, ethambutol, tetracycline and chloramphenicol, suggesting that a *mar*-like system is present in mycobacteria [129]. Following this, Zhang et al. showed that in *M. smegmatis*, Ms6508 is a MarR-like protein whose corresponding *marRAB* operon confers rifampin resistance [130]. However, both of these studies were performed in the nonpathogenic organism *M. smegmatis*. To our knowledge, only one other study has focused on the *mar* operon in *M. tuberculosis*. Radhakrishnan et al. showed that *Rv0678* is a MarR-like regulator that controls transcription of

the MmpS5-MmpL5 transporter, providing direct evidence of a *mar*-like operon in *M. tuberculosis* [131].

Despite the function of its homologs in drug resistance, loss of function mutations in *Rv2887* do not confer resistance to rifampin or isoniazid. However, intact *Rv2887* does negatively regulate the transcription of *Rv0558, Rv0559c, Rv0560c* and *Rv2463*. *Rv2463*, also known as lipP, is a lipase that has some homology to *E. coli* MarA, the activator in the *mar* system, which in *E. coli* is negatively regulated by MarR. While the only known function of *Rv0559c* is that it is secreted, *Rv0560c* methylates benzoquinone (coenzyme Q) and *Rv0558* methylates menaquinone, both of which play an important role in electron transport. Thus, mutation of *Rv2887* may confer resistance to MP-III-71 by up-regulating expression of these two genes, resulting in altered cellular energy levels and transport. This may lead to methylation of MP-III-71, causing inactivation and resistance.

This study revealed that *M. tuberculosis* susceptibility to MP-III-71 is *Rv2887*-dependent. *Rv2887* is a MarR-like protein which inhibits the expression of at least six genes, including *Rv0558*, a menaquinone methyltransferase, and *Rv0590c*, a benzoquinone methyltransferase. A methylated derivative of MP-III-71 is inactive. This suggests that loss of *Rv2887* leading to elevated methylase expression accounts for MP-III-71 resistance in *M. tuberculosis Rv2887* mutants.

4.5 Figures

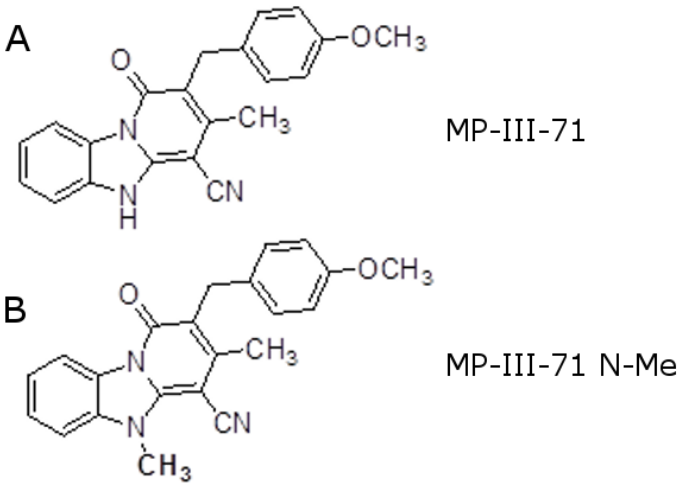


Figure 4-1. Novel compounds used in this study.

Chemical structure of (A) MP-III-71 and (B) MP-III-71-N-Me.

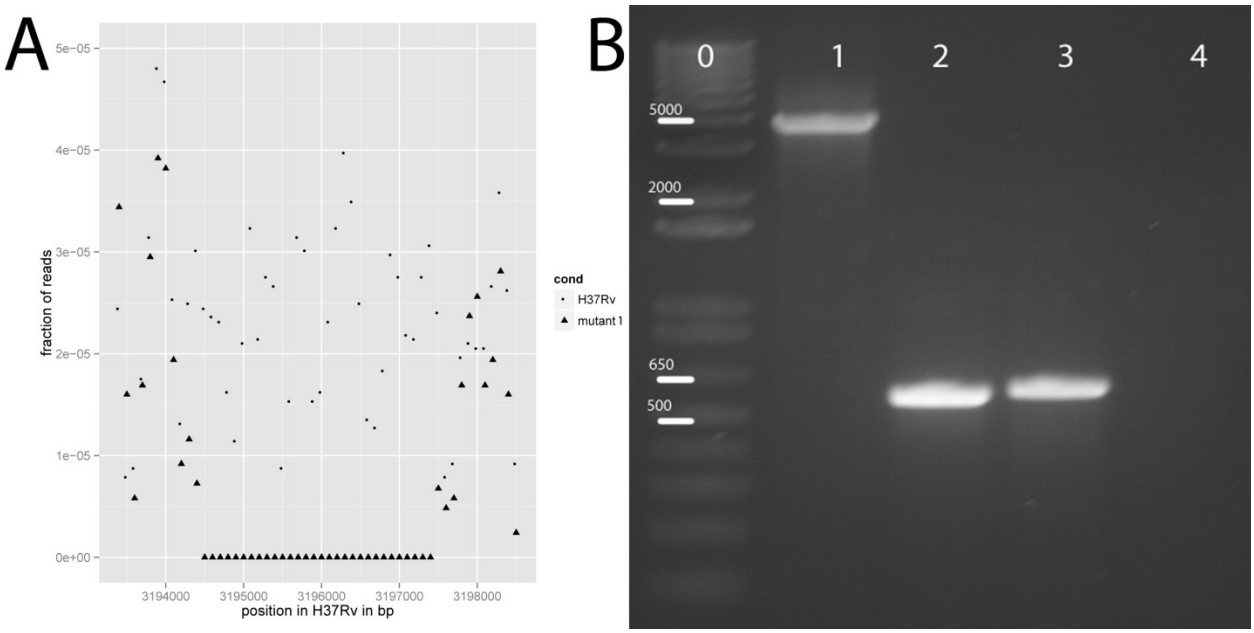


Figure 4-2. 3,183 bp deletion in mutant 1, which includes *Rv2887*.

(A) Coverage analysis of sequencing results for *M. tuberculosis* H37Rv and mutant 1, showing lack of reads in this region for the mutant but not the parent H37Rv strain. (B) Confirmation of the deletion mutation. Lane 0 is the Invitrogen 1kb Plus DNA Ladder. Numbers over the bars indicate the size in base-pairs of the ladder. Lane 1 is a PCR of genomic DNA from the parent H37Rv strain with primers outside the region of the deletion. Lane 2 is a PCR of genomic DNA from mutant 1 with the same primers as lane 1. Lane 3 is a PCR of genomic DNA from H37Rv with primers inside the deletion (in *Rv2887*). Lane 4 is a PCR of mutant 1 genomic DNA with the same primers as lane 3.

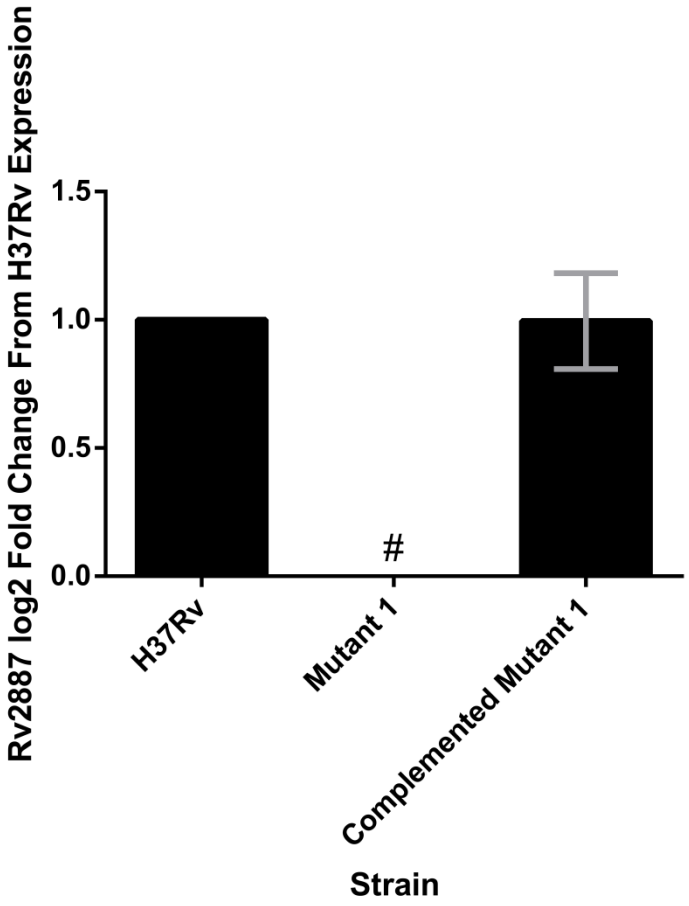


Figure 4-3. Expression of *Rv2887*.

Quantitative reverse transcription PCR of *Rv2887*, normalized to *sigA*. This shows *Rv2887* expression is abolished in mutant 1 but restored to wild-type levels in the complement.

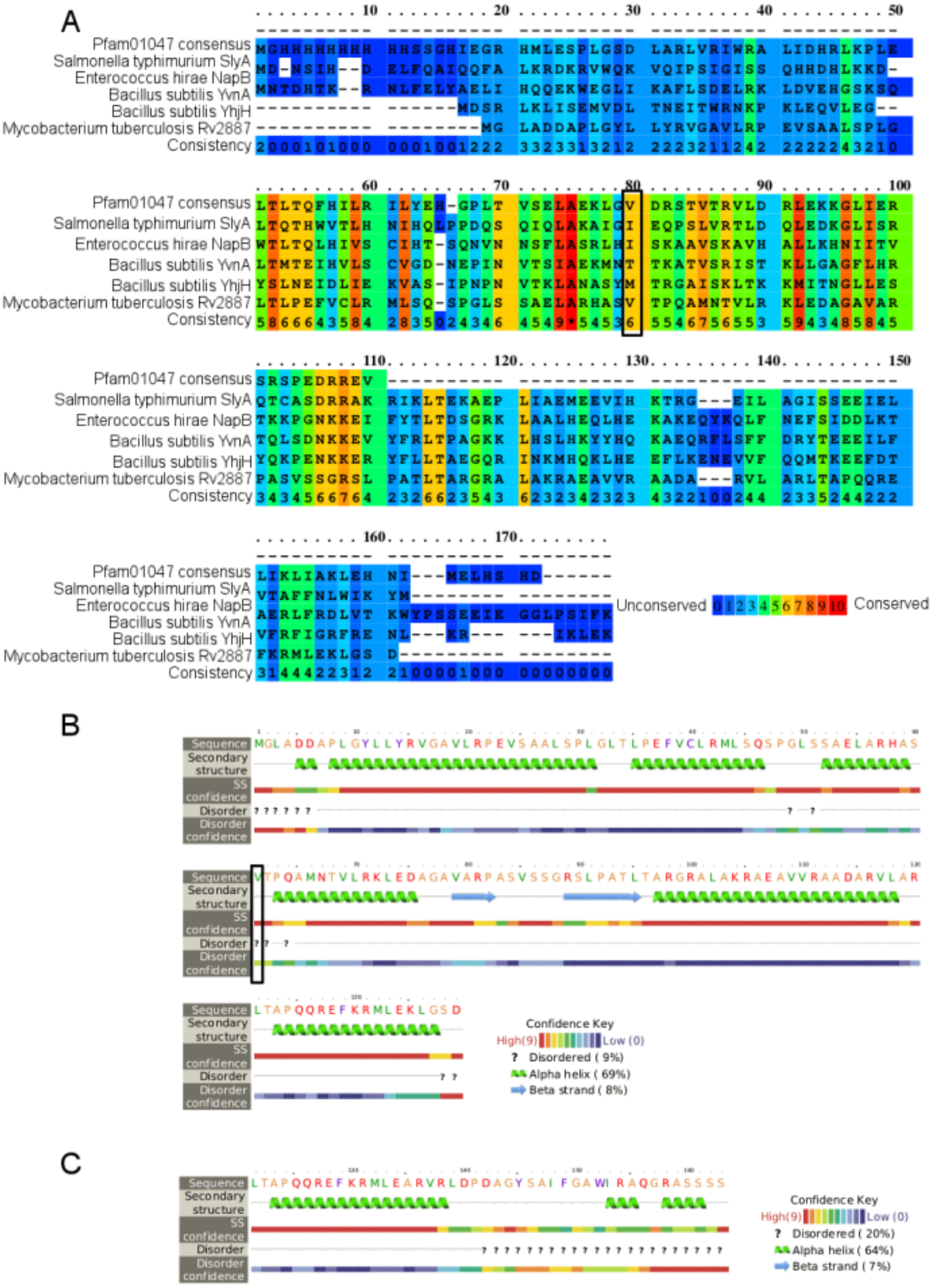


Figure 4-4. Rv2887 is in the MarR family.

(A) Alignment of Rv2887 to the consensus sequence for Pfam01047 (MarR) and four other MarR family proteins (selected as the five most diverse members in the conserved domain database, including the consensus sequence) [95]. The bottom row is the consistency, as estimated by PRALINE [98]. (B) Secondary structure prediction from Phyre2 for Rv2887[97]. (C) Secondary structure prediction for the C terminus of mutant 3. The black boxes in A and B indicate the V61, which is mutated to alanine in mutant 2.

4.6 Tables

Whole Genome Sequencing							
	Sample	Accession Number	Sequencing Platform	Number of reads	Average read length (bp)	Average coverage	Purpose
	H37Rv	SRS889709	Ion Torrent PGM	1118469	118	23.90x	WGS
	mutant 1	SRS891424	Ion Torrent PGM	1128556	113	20.30x	WGS
	mutant 2	SRS891455	Ion Torrent PGM	2085715	101	33.90x	WGS
	H37Rv	SRS891666	SOLiD	25731984	50	266.14x	WGS
	mutant 3	SRS893802	SOLiD	24748378	50	264.32x	WGS
RNA-seq							
	Sample	Accession Number	Sequencing Platform	Number of reads	Average read length (bp)		Purpose
Replicate 1	mutant 1 + MP-III-71	SRS891491	Ion Torrent PGM	962352	119		RNA-seq
	mutant 1 + DMSO	SRS891670	Ion Torrent PGM	999721	120		RNA-seq
	mutant 1 complement+ MP-III-71	SRS891672	Ion Torrent PGM	973824	124		RNA-seq
	mutant 1 complement+ DMSO	SRS891793	Ion Torrent PGM	929050	130		RNA-seq
Replicate 2	mutant 1 + MP-III-71	SRS891794	Ion Torrent PGM	2304483	85.5		RNA-seq
	mutant 1 + DMSO	SRS891795	Ion Torrent PGM	1027257	99		RNA-seq
	mutant 1 complement+ MP-III-71	SRS891796	Ion Torrent PGM	1657712	94		RNA-seq
	mutant 1 complement+ DMSO	SRS891857	Ion Torrent PGM	1229293	199		RNA-seq
Replicate 3	mutant 1 + MP-III-71	SRS891859	Ion Torrent PGM	899239	89		RNA-seq
	mutant 1 + DMSO	SRS891860	Ion Torrent PGM	1057459	120		RNA-seq
	mutant 1 complement+ MP-III-71	SRS891862	Ion Torrent PGM	681799	103		RNA-seq
	mutant 1 complement+ DMSO	SRS891863	Ion Torrent PGM	1448013	101		RNA-seq

Table 4-1. Next generation sequencing metrics.

Statistics from all sequencing runs. WGS = whole genome sequencing.

Strain	MIC for MP-III-71 (μg/mL)	Complement MP-III-71 MIC (μg/mL)	Sequencing Platform	Rv2887 Mutation
H37Rv	0.25-0.5		SOLiD and PGM	
Mutant 1	2	0.25-0.5	PGM	Deletion from 3194362-3197545 (3183 bp), which includes Rv2885c, Rv2886c, Rv2887, and Rv2888c
Mutant 2	1-2	0.25-0.5	PGM	2bp deletion at 3196830-1, which is nucleotide 399 in Rv2887 (amino acid 133)
Mutant 3	1	0.25-0.5	SOLiD	V61A non-synomous SNP in Rv2887

Table 4-2. Summary of Mutants.

SNPs					
Mutation Position in H37Rv	Reference Base	Observed Base	Mutation Annotation	Mutation Effect	Strain Containing Mutation
342542	A	G	Rv0282	D138G	mutant 1 and mutant 2
666418	G	A	Rv0573c	L275L	mutant 1 and mutant 2
2170598	C	T	Rv1918c	V5V	mutant 1 and mutant 2
2654064	A	G	Rv2374c	X344R	mutant 1 and mutant 2
3894771	A	G	Rv3478	T116A	mutant 1 and mutant 2
3894773	C	T	Rv3478	T116T	mutant 1 and mutant 2
4374865	C	A	Rv3892c	V273V	mutant 1 and mutant 2
2979619	A	G	Rv2658c	W24R	mutant 2
4110523	C	T	Rv3668c	intergenic	mutant 2
3196612	T	C	Rv2887	V61A	mutant 3
INDELs					
Mutation Position in H37Rv	Reference Base	Observed Base	Mutation Annotation	Mutation Location	Strain Containing Mutation
55533	TGCC	T	Rv0050	CDS	mutant 1 and mutant 2
131174	T	TG	Rv0108c	Intergenic	mutant 1 and mutant 2
234496	C	CGT	Rv0197	CDS	mutant 1 and mutant 2
424320	T	TC	Rv0354c	CDS	mutant 1 and mutant 2
1274805	CG	C	Rv1146	Intergenic	mutant 1
1780586	C	CG	Rv1575	CDS	mutant 1
3580636	CT	C	Rv3203	Intergenic	mutant 1
3590686	G	GC	Rv3213c	Intergenic	mutant 1
3635239	GC	G	Rv3255c	CDS	mutant 1
3862472	GA	G	Rv3443c	Intergenic	mutant 1
4232224	AGGTGGAGAC AGTGGGCT	A	Rv3785	CDS	mutant 1
3196829	GGA	G	Rv2887	intergenic	mutant 2

Table 4-3. SNPs and indels identified in whole genome sequencing not present in parent H37Rv strain.

Strain	Gene Mutated	length of protein	transposon point of insertion	Reason tested	MIC MP-III-71
CDC1551	NA				0.125-0.25 ug/mL
JHU2658c-26	Rv2658c	369	26	Mutated in mutant sequence	0.125 ug/mL
JHU2658c-30	Rv2658c	369	30	Mutated in mutant sequence	0.125 ug/mL
JHU3668c-441	Rv3668c	699	441	Mutated in mutant sequence	0.25-0.125 ug/mL
JHU0559c-124	Rv0559c	339	124	RNA-seq hit	0.25 ug/mL
JHU0559c-202	Rv0559c	339	202	RNA-seq hit	0.25 ug/mL
JHU2463-1183	Rv2463	1185	1183	RNA-seq hit	0.25 ug/mL

Table 4-4. MIC of MP-III-71 against select transposon mutants.

MICs are given in µg/mL.

Drug	H37Rv	mutant 1	mutant 1 complement
MP-III-71	0.125-0.5	1-2	0.125-0.5
Me-MP-III-71	>32	>32	>32
Isoniazid	0.04	0.04	0.04
Rifampin	0.125	0.125	0.125
Ethambutol	1-2	1-2	1-2
Kanamycin	2	2	2
Verapamil	200	200	200
MP-III-71 + 50 ug/mL Verapamil	0.0625	1	0.3125
Chlorpromazine	10	10	10
Salicylate	250-500	250-500	250-500
MP-III-71 + 62.5 ug/mL Salicylate	0.125	0.25	0.125
CCCP	8	8	8
MP-III-71 + 1 ug/mL CCCP	0.0625-0.5	0.5	0.125-0.5

Table 4-5. MIC of select compounds against H37Rv, mutant 1 and mutant 1 complement.

MICs are given in µg/mL. CCCP=carbonyl cyanide 3-chlorophylhydrazone.

	mutant 1 + MP-III-71 vs. complement + MP-III-71		mutant 1 + DMSO vs. complement + DMSO		
Gene	**Log_2fold change**	**Adjusted p-value**	**Log_2fold change**	**Adjusted p-value**	**Tuberculist annotation**
Rv0560c	-7.67	6.98E-58	0.00	3.29E-53	Possible benzoquinone methyltransferase (methylase)
Rv2887	7.16	4.38E-20	75.40	1.38E-12	
Rv0558	-3.06	1.16E-12	0.08	1.96E-14	menH: Probable ubiquinone/menaquinone biosynthesis methyltransferase MenH
Rv0559c	-3.63	4.70E-11	0.05	3.38E-13	Possible conserved secreted protein
Rv2463	-1.59	0.01	0.35	0.02	lipP: Probable esterase/lipase
Rv2886c	2.34	0.03	3.28	1.00	Probable resolvase

Table 4-6. Significant hits from RNA-seq analysis.

	Hit (Subject)	% identity	Alignment length	Number of mismatches	Number of gap starts	Query start	Query end	Subject Start	Subject end	Expect value	Bit score
MarR (gi|13384163) as query	Rv1404	34.67	75	49	0	37	111	42	116	4.00E-07	47.4
	Rv0880	27.27	110	69	2	4	113	9	107	3.00E-05	41.2
	Rv2887	25.38	130	96	1	13	142	8	136	2.00E-04	38.1
	Rv0737	27.16	81	57	1	33	111	49	129	3.00E-04	37.7
	Rv0042c	32.39	71	48	0	37	107	82	152	0.027	31.2
	Rv2922c	33.33	45	30	0	97	141	723	767	2.4	24.6
	Rv1442	33.33	24	16	0	72	95	49	72	3.9	24.3
	Rv1663	28.57	35	25	0	103	137	322	356	4.6	23.9
MarA (gi|190905560) as query	Rv1931c	30.49	82	55	2	28	108	175	255	4.00E-09	53.5
	Rv3833	26.51	83	60	1	18	100	156	237	1.00E-05	42
	Rv3736	28.92	83	56	2	27	108	246	326	0.017	31.6
	Rv2463	44.83	29	12	1	42	70	168	192	0.23	27.7
	Rv0543c	30.23	43	30	0	7	49	55	97	1.7	25
	Rv0672	25.4	63	38	4	24	86	22	75	6.2	23.1

Table 4-7. Results of using BLAST to compare *E. coli* MarR and MarA to all H37Rv proteins.

Chapter 5: Whole genome sequencing of *Mycobacterium africanum* strains from Mali provides insights into the mechanisms of geographic restriction

Data analysis was performed with the assistance of Dr. Abigail McGuire.

5.1 Abstract

Mycobacterium africanum, made up of lineages 5 and 6 within the *Mycobacterium tuberculosis* complex (MTC), causes up to half of all tuberculosis cases in West Africa, but is rarely found outside of this region. The reasons for this geographical restriction remain unknown. Possible reasons include a geographically restricted animal reservoir, a unique preference for hosts of West African ethnicity, and an inability to compete with other lineages outside of West Africa. These latter two hypotheses could be caused by loss of fitness or altered interactions with the host immune system. We sequenced 92 MTC clinical isolates from Mali, including two lineage 5 and 24 lineage 6 strains. Our genome sequencing assembly, alignment, phylogeny and average nucleotide identity analyses enabled us to identify features that typify lineages 5 and 6 and made clear that these lineages do not constitute a distinct species within the MTC. We found that in Mali, lineage 6 and lineage 4 strains have similar levels of diversity and evolve drug resistance through similar mechanisms. In the process, we identified a novel streptomycin resistance mutation. In addition, we found evidence of person-to-person transmission of lineage 6 isolates and showed that lineage 6 is not enriched for mutations in virulence-associated genes, suggesting lineage 6 does not have an overall loss in virulence. This is the largest collection of lineage 5 and 6 whole genome sequences to date, and our assembly and alignment data provide valuable insights into what distinguishes them from other MTC lineages. These lineages do not appear to be geographically restricted due to an inability to

transmit between West African hosts or an overall loss of virulence. However, lineage-specific mutations, such as mutations in cell wall structure and secretion systems, provide alternative mechanisms that may lead to host specificity.

5.2 Introduction

Mycobacterium africanum is a member of the *Mycobacterium tuberculosis* complex (MTC) that causes up to half of all tuberculosis cases in West Africa [132]. First identified by Castets in 1968, it was originally characterized as having biochemical characteristics intermediate between *Mycobacterium tuberculosis*, which consists of lineages 1, 2, 3, 4, and 7 and is the main cause of human tuberculosis, and *Mycobacterium bovis*, an animal-adapted lineage which causes bovine tuberculosis [133]. Later work divided *M. africanum* into two lineages, *M. africanum* West African type I and *M. africanum* West African type II, which became known as lineages 5 and 6, respectively, within the MTC [134,135].

Lineages 5 and 6 cause a disease similar to classically defined *M. tuberculosis*, although it has been suggested that human disease caused by these lineages may differ compared to that caused by lineages 1-4. For example, patients with lineage 6 disease have been reported to show attenuated ESAT-6 responses compared to patients with classical *M. tuberculosis* lineage disease [136,137]. In addition, in liquid culture systems it has been reported that *M. africanum* has a slower growth rate with a larger bacillary size than *M. tuberculosis* [138,139]. While some studies have found that *M. africanum* is less virulent than *M. tuberculosis*, both in animal models and human patients [138,140-142], others show that there is no difference [143]. Though these contradicting results may be due to differences in the study populations, they underscore how little is known about lineages 5 and 6.

Contributing to our overall lack of knowledge, lineages 5 and 6 are not widely distributed around the globe, unlike lineages 1, 2, 3 and 4 [2]. They are found almost exclusively in patients living in West Africa, with very few cases occurring outside of this region, mostly involving recent immigrants from West Africa [132]. The reason for this apparent geographic restriction is unknown. One hypothesis is the presence of an undiscovered animal reservoir endemic to West Africa, which is supported by the close relationship between lineages 5 and 6 and the animal-adapted lineages of the MTC [144,145]. Another hypothesis is that lineages 5 and 6 have a unique predilection for humans with genetic backgrounds common in West Africa. In fact, using a retrospective epidemiological study of MTC in Ghana, Asante-Poku et al. showed that lineage 5 is associated with the Ewe ethnic group [146]. A third hypothesis is that lineages 5 and 6 are unable to compete with other lineages outside of West Africa, either due to loss of fitness or decreased transmissibility, thus explaining the limited global distribution.

Historically, mycobacterial species were defined by biochemical assays, but, as genetic tools became more readily available, it is now possible to identify genomic regions that define MTC lineages [147]. The publication of the whole genome sequence of *M. africanum* GM041182, a single lineage 6 strain, provided valuable insights into the genetics of this lineage [148]. For instance, the authors identified lineage 6-specific pseudogenes, a novel region not present in *M. tuberculosis*, and single nucleotide polymorphisms (SNPs) in key genes, all of which may play a role in the geographic restriction of lineage 6. A later study sequenced four additional lineage 6 isolates and was able to confirm many of these findings, but also showed that not all mutations identified in *M. africanum* GM041182 are shared by other members of this lineage [139]. To our knowledge, no study has closely analyzed the genetics of lineage 5.

From these studies, it is clear that more sequenced isolates are needed to fully characterize the genetics of lineages 5 and 6 and to illuminate mechanisms that may explain its geographic isolation. Toward this end, we sequenced 92 clinical MTC isolates from Mali, a country in West Africa in which 26.2% and 1.6% of tuberculosis cases are caused by lineage 6 and lineage 5, respectively [149] [132]. Using these and previously published data, we performed both alignment- and assembly-based comparative analyses to further refine our understanding of lineage-specific genomic features that might explain the geographic distribution of lineages 5 and 6. To our knowledge, this is the largest collection of lineage 6 strains sequenced to date, and the first in depth whole genomic characterization of lineage 5.

5.3 Materials and Methods

5.3.1 Samples

101 Mali strains were selected from clinical isolates collected in Bamako, Mali [149], and included all strains identified by spoligotyping as *M. africanum*, *M. tuberculosis* T1, or *M. bovis*. Of these strains, 92 were still viable and were submitted for whole genome sequencing. These 92 strains will be referred to as the "Mali collection". In addition, to improve MTC lineage representation, we selected additional whole genome assemblies that matched the quality of our assemblies. These included four finished *M. bovis* genomes available from GenBank (*M. bovis* AF2122/97 [150], *M. bovis* BCG Mexico [151], *M. bovis* BCG Pasteur 1173P2 [152], and *M. bovis* BCG Tokyo 172 [153]), a set of 40 *M. tuberculosis* strains (9 lineage 1 strains, 12 lineage 2 strains , 7 lineage 3 strains, and 12 lineage 4 strains) from South Africa [154], the finished *M. africanum* genome from Genbank (*M. africanum* GM041182) [148], and *M. canettii* CIPT 140010059 [155]. Combined with the Mali collection, these 137 strains will be referred to as the "assembly collection". Finally, all 161 strains (122 lineage 2, two lineage 3, and 37 lineage 4)

from a study in China were included in the variant analysis to improve geographical and lineage representation [156]. The samples from the China study combined with the samples from South Africa and Mali (for a total of 289 strains) will be referred to as the "alignment collection". Supplemental Tables 5-S1 through 5-S3 lists all strains utilized in this study.

5.3.2 Drug susceptibility testing

Drug resistance to isoniazid, rifampicin, ethambutol and streptomycin was tested for all Mali strains as previously described [149]. We confirmed those results by submitting 17 strains to National Jewish Health in Colorado for agar proportion testing of isoniazid, rifampicin, ethambutol, ofloxacin, niacin, kanamycin, ethionamide, capreomycin, amikacin, cycloserine and para-aminosalicylic acid, as well as radiometric testing of ciprofloxacin and pyrazinamide. The agar proportion results confirmed the mycobacterial growth indicator tube (MGIT) tests performed in Mali. Genotypic drug resistance was determined for rifampicin, isoniazid, ethambutol, streptomycin, ofloxacin, kanamycin and ethionamide using genetic markers from line-probe assays (Supplemental Table 5-S4).

5.3.3 Genome sequencing

Extraction of genomic DNA was performed on 10mL cultures grown in 7H9 broth using the CTAB-lysozyme method as previously described [85]). Library preparation and whole genome sequencing (WGS) were performed as previously described [157-159]. GenBank accessions for all strains used in this analysis can be found in Supplemental Tables 5-S1 to 5-S3, along with assembly statistics for the new sequences generated at the Broad Institute (92 sequences from Mali generated for this study and 40 sequences from K-RITH).

5.3.4 Annotation

All genomes in our assembly collection were uniformly annotated by transferring annotations from *M. tuberculosis* H37Rv. The reference *M. tuberculosis* H37Rv genome (accession CP003248.2) was aligned to draft assemblies using Nucmer [160]. This alignment was used to map reference genes over to the target genomes. Using this methodology, annotations were successfully transferred onto all 137 strains for 3466 of the *M. tuberculosis* H37Rv genes; the rest of the *M. tuberculosis* H37Rv genes transferred to a subset of the genomes.

For those genes not cleanly mapping to *M. tuberculosis* H37Rv, the protein-coding genes were predicted with the software tool Prodigal [161]. tRNAs were identified by tRNAscan-SE [162] and rRNA genes were predicted using RNAmmer [163]. Gene product names were assigned based on top blast hits against SwissProt protein database (>=70% identity and >=70% query coverage), and protein family profile search against the TIGRfam hmmer equivalogs. Additional annotation analyses performed include Pfam [164], TIGRfam [165], KEGG [166], COG [167], GO [168], EC [169], SignalP [170], and TMHMM [171].

5.3.5 Orthogroup clustering and Phylogenetic trees

SYNERGY2 [172-174], available at http://sourceforge.net/projects/synergytwo/, was used to identify cluster-based orthogroups across our assembly collection of 137 genomes. In addition, for each *M. tuberculosis* H37Rv gene, we defined a second set of annotation transfer-based ortholog groups as the set of genes for which annotations were transferred from this *M. tuberculosis* H37Rv gene in our annotation protocol. Genes without *M. tuberculosis* H37Rv orthologs were manually examined in the context of their SYNERGY orthogroups to identify lineage-specific novel genes.

Phylogenetic trees were generated by applying RAxML [175] to a concatenated alignment of 3343 single-copy core SYNERGY cluster-based orthogroups (excluding orthogroups with paralogs) across all 137 organisms. Bootstrapping was performed using RAxML's rapid bootstrapping algorithm (1000 iterations).

5.3.6 Average Nucleotide Identity analysis (ANI)

Calculations of ANI were done using our assembly collection as previously described [176,177], using our SYNERGY cluster-based orthogroups, with a threshold for species identity differing by more than 5%.

5.3.7 Gene Content Analysis

PAUP [178] was used to reconstruct gain and loss of *M. tuberculosis* H37Rv orthologs at ancestral nodes of the assembly collection phylogenetic tree using parsimony, using our ortholog groups based on annotation transfer.

In order to analyze changes in gene content, we used a cost matrix with values of 10 for a gene gain, 5 for a gene loss, and 0.2 for an increase or decrease in copy number. We looked for orthologs found within all members of one clade, and absent in other clades. As a further filter, we also required that orthogroups be found in >80% of the clade of interest, and <20% of other strains. We performed this analysis for five key clades: lineage 5, lineage 6, *M. bovis*, the clade including *M. bovis* and lineage 6, and the clade including lineages 5, 6 and *M. bovis*.

In addition, we selected the Pfam gene categories most expanded or reduced in each clade of interest. We determined significance using Fisher's test ($P<0.05$). For each of the clades described above, we compared the strains below this node versus all other strains in our analysis.

5.3.8 Identification of SNPs

For our alignment collection, reads were mapped onto a reference strain of *M. tuberculosis* H37Rv (GenBank accession number CP003248.2) using BWA version 0.5.9.9 [87]. In cases where read coverage of the reference was greater than 200x, reads were down-sampled using Picard [179] prior to mapping. Variants, including both single nucleotide polymorphisms (SNPs) and multi-nucleotide polymorphisms, were identified using Pilon version 1.5 as described [157] and were used to construct phylogenetic trees using FastTree [180].

We defined lineage-specific variants as those with positive predictive value >95%, negative predictive value >95%, true positive rate >95%, true negative rate >95%, and number of true positives >7. Mutations were considered *M. africanum*-specific (lineage 5 and 6-specific, identified as LIN-Maf in the tables) if they met these cutoffs for lineage 5 and 6 combined and were present in both lineage 5 strains. Similarly, mutations were considered *M. tuberculosis*-specific if they met these cutoffs for lineages 1-4 combined. No *M. tuberculosis*-specific mutations were identified. Due to inclusion of only two lineage 5 strains in our dataset, no lineage-specific variants were identified in lineage 5. Thus, for this lineage only, we used a less stringent requirement to define lineage-specific variants: we required that variants be present in both lineage 5 strains and in <5% of the strains in each other lineage. We classified each gene containing a lineage-specific variant into functional group categories, including Gene Ontology (GO) [168], KEGG [166], Pfam [164], and COG [167]. We then evaluated enrichment using Fisher's Exact test and corrected for multiple comparisons using the Storey method for functional group categories [181].

5.3.9 Identification of pseudogenes

A pseudogene was defined as any gene that had a loss of function mutation anywhere within the coding sequence. Loss of function mutations were defined as nonsense mutations or insertions or deletions that were not in multiples of 3 base pairs or were greater than 30 base pairs. Lineage-specific pseudogenes were determined using the same definitions as for variants on a per gene basis (positive predictive value > 95%, negative predictive value >95%, true positive rate >95%, true negative rate >95%, number of true positives >7, with the exception of lineage 5, which used the SNP cutoffs of pseudogene in both lineage 5 strains and in >5% in each other lineage).

5.3.10 Computational gene function assessments

The effect of select non-synonymous mutations on protein function was assessed using the online version of SIFT at default settings [101], unless there was low confidence in the prediction, in which case SIFT was run for each of the four available databases (UniRef90 from April 2011 [default], UniProt-SwissProt 57.15 from April 2011, UniProt-TrEMBL from March 2009 and NCBI nonredundant from March 2011). Peptide binding was predicted using the NetMHCII online tool with default settings [182].

5.4 Results

5.4.1 *M. africanum* and *M. tuberculosis* lineages are part of the same species

Our Mali collection of 92 clinical MTC strains was isolated from patients presenting with pulmonary tuberculosis at Point G, Bamako, Mali between 2006 and 2010 as part of a cross-sectional study to analyze the diversity of the MTC in Mali [149]. All patients were Mali natives,

with the exception of one patient born in central Africa (Supplemental Table 5-S1). We sequenced this collection using the Illumina sequencing platform, and the resulting reads were both assembled into contigs and aligned against the *M. tuberculosis* H37Rv reference genome. Based on our phylogenetic reconstructions, our collection included one lineage 1, two lineage 2, zero lineage 3, 63 lineage 4, two lineage 5 and twenty-four lineage 6 strains (Figure 5-1). In order to perform statistical comparisons of the *M. tuberculosis*, *M. africanum* and *M. bovis* lineages, this dataset (the Mali collection) was combined with data from additional strains from GenBank and South Africa (assembly collection, Figure 5-2), as well as data from China (alignment collection; see Materials and Methods and Supplemental Tables 5-S1 through 5-S3). These additional comparator genomes enabled us to examine the distinguishing characteristics of lineages 5 and 6 in detail.

Since this represents the largest collection of whole genome sequences of lineage 5 and 6 strains to date, we used our assembly collection to conduct a detailed examination of their phylogeny and characteristics in relation to other members of the MTC, including *M. bovis* and *M. tuberculosis*. *M. bovis* is considered an animal strain that mainly infects cattle and rarely humans, while *M. tuberculosis* is human adapted, and lineages 5 and 6 are thought to be intermediate between the two [132,144]. Using our assembly collection, we constructed a high-resolution phylogenetic tree using 3,343 single-copy core orthogroups (sets of orthologs) conserved across all 137 strains (Materials and Methods). This tree was rooted using the outgroup *M. canettii* and agreed with phylogenies observed by other groups, including the fact that each of the lineages was clearly separated from the other, with lineage 5 being more closely related to human-adapted strains and lineage 6 being more closely related to *M. bovis,* although all of these lineages were very closely related (Figure 5-2) [2,144].

It has been previously shown, using average nucleotide identity (ANI) analysis, that separate bacterial species share <65-90% of genes and have no more than 94-95% ANI among shared genes [176,177]. Using gene content and nucleotide variation among shared genes, we examined the genetic distances between strains within the assembly collection to understand how mycobacterial species fit within this framework. We observed that there was little diversity within the lineages analyzed. Strikingly, values from inter-lineage comparisons of *M. tuberculosis, M. bovis*, and *M. africanum* strains overlapped those from intra-lineage comparisons, showing very little separation, with >99% ANI and >94% fraction of shared genes (Figure 5-3) suggesting that these different organisms should not, in fact, be named different species.

In contrast, MTC pairwise comparisons with *M. canettii* revealed a clear separation between the two groups suggesting that they occupy distinct niches (Figure 5-3). *M. canettii* is a smooth tubercle bacilli that causes human tuberculosis in East Africa and is considered an emerging pathogen in some parts of the world, but its natural host(s) and reservoirs remain unknown [183]. Thus, it might be argued, based on these data and the traditional cutoffs set by ANI analysis, that all MTC members should be named the same species, and that even *M. canettii* should be included since pairwise identities with MTC exceeded these thresholds (Figure 5-3). However, as Smith et al. have previously discussed [184] changes in nomenclature can cause confusion in the literature, and so we will continue to refer to *M. africanum*-associated lineages as either lineage 5 or 6 within the MTC.

5.4.2 Lineage 6 is involved in recent person-to-person transmission events and is as diverse as lineage 4 strains in Mali

Despite the fact that lineages 5 and 6 are so closely related to lineages 1-4, as demonstrated by our ANI plots, they are still unique in being geographically restricted compared to these other lineages. One hypothesis for this restriction is that they are less fit and thus unable to compete with other lineages within the MTC. To examine this possibility, we looked within the Mali collection for clues that lineage 5 and 6 strains were undergoing changes in population structure that might suggest that these lineages are slowly dying out. Using alignments of our Mali collection to *M. tuberculosis* H37Rv, we found that the two lineage 5 strains were not closely related. However, within lineage 6, we observed three pairs of strains that were separated by less than 10 SNPs (see Materials and Methods). There were six such clusters within lineage 4. A cutoff of 12 SNPs has previously been used to determine recent transmission [185]. Thus, strains separated by less than 10 SNPs provide evidence of transmission, suggesting that 6 of 24 (25%) of our lineage 6 strains and 13 of 63 (21%) of our lineage 4 strains were involved in recent transmission events, confirming previous observations based on alternative genotyping approaches that there is robust ongoing transmission of lineage 6 within this region [140].

Alignment-based approaches can miss differences in regions not present in the reference (in this case *M. tuberculosis* H37Rv, which is in lineage 4), so, to further address the question of whether lineage 6 is dying or succeeding, we analyzed the diversity within lineage 6 using our assembly collection and compared this diversity to that of the predominant *M. tuberculosis* lineage in this region, lineage 4. Diversity, as measured by ANI of shared genes, was comparable for lineage 6 and lineage 4 strains from Mali (Figure 5-4), with no statistical

difference among diversity values when comparing between the two groups. Although this result does not eliminate the possibility of differing ecologies, such as an animal reservoir for lineage 6, as has previously been hypothesized [145], it does suggest that lineages 4 and 6 in Mali are under similar evolutionary pressures and are responding to them in a comparable manner. These results also show that lineage 6 strains—which are geographically restricted--are adapting to evolutionary pressure by maintaining diversity comparably to lineage 4 strains from the region which are not geographically restricted.

5.4.3 Lineages 5 and 6 are not enriched for mutations in genes associated with virulence

Given the reports of lineages 5 and 6 strains having decreased virulence [138,140-142], we hypothesized that altered virulence may contribute to geographical restriction, either due to changes in host requirements or to a reduction in fitness. To test this hypothesis, we examined lineage-specific pseudogenes (truncated genes) and non-synonymous SNPs in known essential genes, slow growth genes, and genes required for virulence in mice and growth in macrophages to determine whether lineages 5 and 6 had an enrichment of defects in these genes that might contribute to altered virulence [186-188]. Although both lineage 5 and 6 had lineage-specific mutations in these gene categories (Supplemental Table 5-S5) so did other lineages, and the proportion of mutated genes in lineage 6 was not significantly different from that of the other MTC lineages [139] (Figure 5-5). Lineage 4 was not included on this graph because it only had one lineage-specific mutation in an intergenic region when aligned to *M. tuberculosis* H37Rv, which is a member of lineage 4, and lineage 5 was excluded due to low sample size. We performed a similar analysis on the full length of genes encoding known T cell antigens as defined by Comas et al. to explore whether alterations in these genes might be restricting host

specificity, but again we observed no significant difference in the proportion of lineage 6-specific mutations that fell within these genes as compared to lineages 1, 2 and 3 (Figure 5-5) [135]. Similarly, we looked for enrichment in lineage-specific mutations in COG, GO, KEGG, Pfam and TIGRfam gene categories, but found no enrichment in any of these categories, either for pseudogenes or non-synonymous SNPs. This corroborates our observations from ANI that the lineages of the MTC are very similar in their overall genetic composition and suggests that lineage 6 may not be impaired in virulence. However, while the overall number of mutations in virulence genes was not enriched, we identified mutations in these genes that will be discussed below.

5.4.4 Lineage 6 evolves drug resistance through similar mechanisms to other MTC lineages

Studies have shown that lineages 5 and 6 evolve drug resistance less often compared to other MTC lineages, including the study from which these sequenced strains were obtained [149,189]. Thus, one hypothesis for the limited geographic range of lineages 5 and 6 could be decreased fitness relative to strains better able to evolve antibiotic resistance. In this case, we would expect that mutations driving drug resistance in these two lineages would be different from those evolving in more successful lineages. Thus, we analyzed the entire Mali collection for the presence of mutations known to confer drug resistance. The mutations are used in common nucleic acid-based commercial tests [190,191] for the detection of drug resistance [192-198] (Supplemental Table 5-S4). We compared the identified genotypic drug resistance-conferring mutations with our phenotypic characterization of rifampicin, isoniazid, ethambutol and streptomycin, and had high sensitivity and specificity for isoniazid, rifampicin, and ethambutol, but poor sensitivity for streptomycin (Figure 5-1 and Supplemental Figure 5-S1; Table 5-1). Forty

(60%) strains in lineage 1-4 and only four (15%) of the lineage 5 and 6 strains were resistant to at least one of these drugs. In addition, we identified known mutations in genes associated with resistance to drugs that were not phenotypically assessed including ofloxacin, kanamycin, and ethionamide.

None of the phenotypic streptomycin resistance was explained by mutations contained in our list of known mutations (Figure 5-1). However, when we looked more closely at mutations in genes known to play a role in resistance to streptomycin, we observed that 23 of the 35 strains with unexplained streptomycin resistance harbored a non-synonymous point mutation in the *gidB* gene (Supplemental Figure 5-S1d). While SNPs in the *gidB* gene have previously been reported in association with streptomycin resistance, this particular SNP has not heretofore been identified [199]. This mutation, at position 236, changes a leucine to a serine, and was predicted by the SIFT algorithm [101] to affect GidB protein function. Of the remaining 12 strains with unexplained streptomycin mutation, 10 had different mutations in *gidB*, and 2 had mutations in *rpsL*, another gene known to have a role in resistance to streptomycin [200]. Thus, it appears that strains circulating in Mali might have evolved streptomycin resistance via unique changes in *gidB*.

Using the list of mutations in Supplemental Table 5-S4, we found that 25 (38%) of the Mali strains belonging to lineages 1, 2 or 4 could be classified as MDR (multi-drug resistant; resistant to isoniazid and rifampin), and two (3%) could be classified as pre-XDR (pre-extensively drug resistant; resistant to isoniazid, rifampin, plus either ofloxacin or kanamycin). In contrast, three (11%) of the lineage 5 and 6 strains could be classified as MDR, and one (4%) could be classified as pre-XDR. The presence of these pre-XDR strains is of particular concern, as XDR has not been reported in Mali, and testing is not currently performed routinely for second line

antibiotics [1,201]. However, similar resistance-conferring mutations were found between the lineages (Supplemental Figure 5-S1). Thus, although the sample size was small, our results suggest that drug resistance, while less frequent in lineage 6, evolves through acquisition of similar mutations to lineages 2 and 4 in Mali.

5.4.5 Evolutionary history: Nodes A-D

At each node we used a combination of assembly and alignment data to identify distinguishing characteristics of the putative ancestral strain at that node. First, we used the assembly collection to identify orthologs that were either gained or lost at each node (Table 5-2; Materials and Methods). Many of these genes fell into already known regions of difference (RDs), previously identified by genome hybridizations [134,147]. For some RDs, the first and/or last gene in the region was not identified in our analysis because enough of the gene remained to align to *M. tuberculosis* H37Rv, and thus was not considered absent. Second, we used functional Pfam annotation of genes from the assembly collection to detect protein domains that were significantly enriched or reduced within the members of each group. Third, we used variant calls from our alignment collection to identify other mutations (smaller than a gene) that were enriched or specific to each group, including those that caused truncations of genes (pseudogenes) that are likely to affect protein function (Tables 5-2, Supplemental Tables 5-S5 through 5-S8). Our alignment collection enabled us to more accurately determine what features were specific to a particular lineage, and showed that many mutations and pseudogenes previously identified as lineage-specific were not lineage-specific when evaluating our larger set of strains (Table 5-3 and Supplemental Table 5-S8).

5.4.5.1 Node A: root node of lineage 6

The Mali collection provided a large and diverse collection of lineage 6 strains, which allowed us to more fully characterize the genetic content of this lineage than had been done previously [139,148]. We observed no gene gains at the root node of lineage 6 (Node A in Figure 5-2), and only one gene loss (previously identified as part of RD701), as expected from earlier phylogenetic work (Table 5-2) [134].

However, the traits resulting in geographic restriction of this lineage are likely due to smaller variants within genes or in intergenic regions altering gene expression, rather than in large genetic changes. When reads from lineage 6 strains were aligned to *M. tuberculosis* H37Rv, we observed 681 lineage 6-specific mutations, 82 of which were in intergenic regions and the rest (599) were within coding sequences, including 8 that resulted in truncated proteins likely to have abolished function (pseudogenes) (Table 5-4, Supplemental Tables 5-S5 to 5-S7). Five of these pseudogenes had previously been identified, while three were novel (Table 5-3 and Supplemental Table 5-S8) [148]. Table 5-S8 provides a side-by-side perspective of the lineage 6 pseudogenes that we detected and demonstrates that the genes identified by Bentley et al., but not by this study, were either pseudogenes in some other lineages, or were not pseudogenes in all lineage 6 strains.

5.4.5.2 Node B: root node of lineage 6 and M. bovis

M. africanum was first identified due to its intermediate phenotype between *M. bovis* and *M. tuberculosis*, which was confirmed by its genetic position on the MTC tree [133,144]. Based on this tree, and the tree generated by our own data, lineage 6 and *M. bovis* share a root ancestor separate from all other MTC lineages (Node B in Figure 5-2). Thus, although we did not have alignment data for *M. bovis,* we were interested in how the genetic content of these two

lineages compared. *M. bovis* and lineage 6 shared loss of RD10, RD7, and RD8 compared to lineage 5 (Table 5-2), corresponding with previous observations [147]. Our Pfam analysis revealed that these two lineages also share a copy number reduction in the MCE Pfam category (PF02470.15), as a result of loss of MCE operon 3, which is contained in RD7. In addition, we identified two genes that were gained at this node (Table 5-2). One of these genes is a PE-PGRS protein; the other is a hypothetical protein.

5.4.5.3 Node C: root node of lineage 5

To our knowledge, this is the first study to perform an in-depth analysis of the whole genome sequence of lineage 5 isolates. Though our dataset contained only two lineage 5 strains, they provided valuable insights into the specific features of this lineage. For example, our gene content analysis revealed the loss of RD711, RD713, and RD743 at Node C, as expected, but also identified two additional lineage specific genes losses that were not part of any known region of difference [134] (Table 5-2). One of these genes is *Rv1523*, which is annotated as a methyltransferase, and is most likely S-adenosyl-L-methionine dependent. KEGG predicts this gene to be in the pathways for tyrosine metabolism (ko00350) and polycyclic aromatic hydrocarbon degradation (ko00624). The other novel lost gene is *Rv3514*, annotated as PE-PGRS57. This is part of a larger family of proteins that are highly polymorphic and may play a role in antigenicity.

In addition to these larger changes, we identified 952 lineage 5-specific mutations and 43 lineage 5- specific pseudogenes (Table 5-4, Supplemental Tables 5-S5 to 5-S7). The larger number of lineage-specific mutations and pseudogenes compared to the other lineages was a result of our small sample size, which required us to use different cutoffs. However, the average number of mutations per strain for lineage 5 was comparable to the other lineages in our study

(Table 5-4). Nevertheless, given the lack of literature on this lineage, these mutations provide novel insights into the distinguishing characteristics of this lineage.

5.4.5.4 Node D: root node of lineage 5, 6 and M. bovis

Through genomic hybridization studies, it is known that all *M. bovis* and *M. africanum* strains share the loss of RD9 [147]. Our gene content analysis of Node D confirmed this finding and identified one additional gene lost at this node and one gained gene (Table 5-2). The lost gene was *Rv2084*, which is annotated as a hypothetical protein but had BLAST similarity to a TetR family transcriptional regulator. The gained gene was a PPE family protein. In addition, our analysis of changes in Pfam content revealed loss of two families. One of these was PF13276.1, which consists of proteins with a helix-turn-helix domain. The helix-turn-helix is a motif involved in DNA binding. This suggested a change in gene expression in these three lineages, a suggestion supported by the fact that the other Pfam domain lost at this node was PF12349.3, a sterol-sensing Pfam domain. Since we did not have alignments of any *M. bovis* strains, we focused on mutations shared between lineages 5 and 6. There were 90 shared mutations, including 5 pseudogenes (Tables 5-S5 to 5-S7).

5.4.6 Individual lineage-specific features suggest additional mechanisms that could be involved in geographic restriction

5.4.6.1 Mutations in ESX secretion systems are common in all MTC lineages

One distinguishing clinical characteristic of lineage 6 is an attenuated T cell response to ESAT-6 in patients infected with this lineage as compared to patients infected with lineages 1-4 [136]. This altered immune response supports the hypothesis that there is lineage 5 and 6 specificity for a particular host immunogenic background. While it was hypothesized that the

attenuated immune response was due to mutation of *Rv3879c*, which is part of the ESX-1 secretion pathway, complementation of this gene did not restore ESAT-6 secretion or presentation to T cells, suggesting that defective *Rv3879c* is not the cause of the altered immune response [138]. Although Bentley et al. reported that *Rv3879c* was a pseudogene in *M. africanum* GM041182, we found that the pseudogenization of *Rv3879c* was not lineage specific, as not all lineage 6 strains encoded truncated versions of this protein, and many strains in other lineages were also truncated (Supplemental Table 5-S8)[148]. This suggested that inactivation of *Rv3879c* is unlikely to explain either the host preference or the altered host immune response of lineages 5 and 6. In fact, *Rv3879c* had lineage specific mutations in lineages 1, 2 and 5, but not lineage 6 (Table 5-5, Supplemental Tables 5-S5, 5-S9). However, we did observe lineage 5 and 6 specific polymorphisms in other genes involved in ESX secretion systems that might explain the different immune responses of lineage 6-infected patients as compared to those infected with other lineages (Table 5-5). Eight of the nonsynonymous mutations were predicted to affect protein function, but even mutations not predicted to affect protein function might also affect the antigenicity of these secretion systems [101]. In fact, we observed lineage specific mutations in ESX-encoding genes in all lineages, suggesting that each lineage may have unique interactions with the host (Table 5-5). Thus, our data show that ESX-secretion system polymorphisms are common across all MTC lineages and are not unique to lineages 5 and 6.

5.4.6.2 Alterations in abundantly secreted proteins in lineages 5 and 6

We also observed lineage 6 specific polymorphisms in other genes that are predicted to have a role in modulating the host's immune response. For example, we identified nonsynonymous changes in the gene encoding antigen 85B (*Rv1866*), a secreted immunogenic protein that has been proposed as a potential vaccine target [202] (Table 5-5, Supplemental

Tables 5-S5 and 5-S9). No other lineages had a lineage-specific mutation in this gene. Though SIFT predicted that this change would not affect protein function, NetMHCII, an online tool that predicts binding of peptides to MHC class II alleles, predicted that that this mutation would change the binding of one of the putative strong binding peptides [101,182]. Although lineage 5 did not contain lineage-specific mutations in antigen 85B, we detected a lineage 5 specific SNP in another abundantly secreted protein, MPT64 (*Rv1980c)* (Table 5-5, Supplemental Tables 5-S5 and 5-S9). This mutation was predicted to affect both protein function and binding to MHC class II molecules [101,182]. Thus, part of the reason for geographical restriction of lineages 5 and 6 may be due to alterations in the pathogen-host immune interaction. This is a particular concern for vaccine development in West Africa.

5.4.6.3 Mutations in L,D transpeptidases

It has previously been shown that *M. africanum* GM041182 has a distinct physiology as compared to of *M. tuberculosis* H37Rv, as *M. africanum* GM041182 had a larger cell size and slower growth rate [138]. Possibly explaining these differences, we identified lineage 6-specific nonsynonymous SNPs in genes encoding the L,D transpeptidases, *LdtA* and *LdtB* (*Rv0166c* and *Rv2518c*), shown to form cross-linkages within peptidoglycan (Table 5-5, Supplemental Tables 5-S5 and 5-S9) [203]. When LdtA and LdtB homologs were inactivated in *M. tuberculosis* CDC1551 (LdtMt1 and LdtMt2), differences in cell shape, size, surface morphology, growth and virulence were observed [204]. Only the SNP in LdtA was predicted to affect protein function [101]. Lineage 5 also contained a non-synonymous SNP predicted to affect protein function in LdtA (Table 5-5, Supplemental Tables 5-S5 and 5-S9). No other lineages had a lineage-specific mutation in an L,D-transpeptidase.

5.4.6.4 Loss of mammalian cell entry (MCE) proteins in lineages 5 and 6

M. tuberculosis H37Rv contains four mammalian cell entry (MCE) operons, which play an important role in mycobacterial virulence [205]. Beyond the earlier report that all lineage 6 strains lacked operon 3 as part of RD7 [147] (Table 5-2), we observed lineage 6-specific nonsynonymous SNPs impacting a protein from MCE operon 1 and two proteins from MCE operon 2, as well as a nonsynonymous mutation in the gene encoding Mce1B shared with lineage 5. In lineage 5 strains, the MCE operons 1 and 3 had nonsynonymous mutations. Only MCE operon 4 did not contain a lineage-specific mutation in either lineage, while operons 1 and 3 were mutated in both (Table 5-5, Supplemental Tables 5-S5 and 5-S9). In *M. tuberculosis* H37Rv, each of these operons has a different expression profile in culture and, when deleted, they have distinct growth defect phenotypes in mice [187,206]. Thus, the presence of a wild-type operon 4 in lineages 5 and 6 may not compensate for mutations in other operons that impact MCE function. In comparison, the other lineages had nearly identical MCE operons as compared to *M. tuberculosis* H37Rv (Table 5-5, Supplemental Tables 5-S5 and 5-S9).

5.4.6.5 Alterations in metabolism in lineages 5 and 6

Lineage 6 had lineage-specific mutations, including pseudogenes, in multiple components of important biosynthetic pathways, such as molybdenum and cobalamin synthesis (Table 5-5, Supplemental Tables 5-S5, 5-S7 and 5-S9). Molybdenum cofactors are key catalysts for redox reactions, and are an important part of the evolution of pathogenic mycobacteria [207]. We detected lineage 6-specific nonsynonymous SNPs in two molybdopterin biosynthesis proteins and in the gene encoding the molybdenum transporter, ModC (Table 5-5, Supplemental Tables 5-S5 and 5-S9). In addition, mycobacteria are one of the few bacterial pathogens with the ability to synthesize vitamin B12, another important cofactor [208]. Bentley

et al. reported that two of the genes encoding the biosynthesis proteins, CobL and CobK, are pseudogenes in all lineage 5 and 6 strains [148]. We found that loss of function in *cobL* is specific to both lineage 5 and 6 but *cobK*, while a pseudogene in both lineages, is also a pseudogene in all lineage 1 strains and some lineage 2 and lineage 4 strains (Table 5-5, Supplemental Table 5-S7 and 5-S8). We also observed a lineage 6 specific nonsynonymous SNP in c*obD*, although SIFT predicted that this change would not affect protein function [101] (Supplemental Table 5-S5). Loss of these cofactor biosynthetic pathways could have ramifications on the function of proteins that use these cofactors, and thus could have indirect effects on host-pathogen interactions.

The vitamin B12 pathway was also mutated in lineage 5. Lineage 5 had a nonsynonymous mutation in four of the cobalamin synthesis enzymes and two flavoproteins. In addition, there was a nonsynonymous mutation in a riboflavin biosynthesis protein, part of the vitamin B3 biosynthetic pathway (Table 5-5, Supplemental Tables 5-S5 and 5-S9). The mutations in *cobO, cobM* and *cobU*, all important parts of the vitamin B12 pathway, were all predicted to affect protein function [101]. Similarly, one of the molybdenum cofactor biosynthesis genes was a pseudogene while another contained a nonsynonymous mutation (Table 5-5, Supplemental Tables 5-S5, 5-S7 and 5-S9). Besides the pseudogene in *cobK* in lineage 1 and some lineage 2 and 4 strains, lineage 2 had an insertion in *cobB* resulting in a pseudogene, while lineage 1 had nonsynonymous mutations in three molybdenum-associated genes (Supplemental Table 5-S7). Likewise, only lineage 1 had a nonsynonmous mutation (a deletion) in a riboflavin-associated gene (Table 5-5, Supplemental Tables 5-S5 and 5-S9). Thus, though mutations in these genes are seen in other MTC lineages, lineages 5 and 6 seem to have an increased number of mutations in cobalamin and vitamin B biosynthesis pathways. Differences in cofactor synthesis could indicate

a in a lineage 5 and 6 preference for a different host environment, as in the case of an animal reservoir or variation in host immune pressures.

5.4.6.6 Mutations in adenylate cyclase

Interestingly, both lineage 5 and lineage 6 had mutations in genes encoding adenylate cyclases, though the affected genes were different between the two lineages. Adenylate cyclase makes cyclic AMP (cAMP), an important cell signaling molecule. Deletion of one of the 17 adenylate cyclases in *M. tuberculosis*, *Rv0386*, reduces virulence and alters the immune response [209]. Bentley et al. found that this gene was a pseudogene in *M. africanum* GM041182. Although this mutation was lineage 6-specific, lineage 2 also had several strains in which *Rv0386* was a pseudogene, suggesting that loss of this particular adenylate cyclase is unlikely to explain the geographical restriction of *M. africanum* (Supplemental Tables 5-S5 and 5-S7). However, lineage 5 had nonsynonymous mutations in two adenylate cyclases, *Rv1320c* and *Rv1647*, both of which were predicted by SIFT to affect protein function [101]. Lineage 6 had a lineage-specific insertion in *Rv1264*, and nonsynonymous mutations in two other adenylate cyclases, although these mutations were tolerated according to SIFT. In contrast, lineages 2 and 3 had no nonsynonymous mutations that were in an adenylate cyclase, while lineage 1 had one synonymous and one nonsynonymous mutation predicted to not affect protein function (Table 5-5). Thus, there may be differences in cAMP signaling within lineages 5 and 6, particularly lineage 5, which could impact how this lineage interacts with the host.

5.4.6.7 Mutations in drug resistance-associated genes

Although we found that lineage 6 evolves drug resistance through similar mechanisms to lineages 2 and 4 (Figs 1 and S1), lineage 6 appears to evolve drug resistance at lower rates, and so we hypothesized that there may be mechanisms that could prevent the development of

resistance. Thus, we screened our lineage-specific SNPs to identify mutations in genes associated with drug resistance [149,189]. Interestingly, lineage 6 had two lineage-specific nonsynonymous mutations in *rpoB*, the gene that confers resistance to rifampicin when mutated in specific regions, and one lineage-specific nonsynonymous mutation in *embC*, a gene that can confer resistance to ethambutol when mutated (Supplemental Figures 5-S1a and 5-S1c, Table 5-5, Supplemental Tables 5-S5 and 5-S9). Lineage 1 and 3 also had lineage-specific mutations in *embC* but no other lineage had a lineage specific nonsynonymous mutation in *rpoB*. Both of the *rpoB* mutations were predicted to be tolerated by SIFT and do not confer drug resistance, but may have an effect on the development of resistance conferring mutations in *rpoB*, helping to explain the decreased rate of MDR in lineage 6 [101]. In addition, lineage 5 had nonsynonymous mutations in genes encoding AtpH (*Rv1307*) and AtpG (*Rv1309*), both of which are subunits of ATP synthase [210] (Supplemental Table 5-S5). Both of these mutations were predicted to affect protein function by SIFT [101]. ATP synthase is a target of bedaquiline, a new antibiotic reserved for the treatment of drug resistant tuberculosis [211].

5.5 Discussion

Our study describes the largest collection of sequenced lineage 6 isolates to date, and, to our knowledge, the first in depth analysis of the genetics of lineage 5. Through our work, we have characterized the genetic basis of antibiotic resistance in lineage 6 strains from Mali, shown that *M. africanum* and *M. tuberculosis* are part of the same species that exhibit similar intra-lineage diversity, and better defined the mutations and changes in gene content that typify these lineages. Collectively, this work provides insights into these understudied lineages and provides hypotheses as to why they are geographically restricted.

We evaluated 92 Mali MTC isolates using both assembly and alignment based approaches. Our assemblies revealed several new regions of difference and our alignments identified smaller lineage-specific changes. In addition, we found that not only is *M. africanum* not its own species, but some *M. africanum*-*M. tuberculosis* pairs of strains are more closely related than some pairs of strains from the same lineage. This emphasizes the extremely close relationship between all MTC lineages, highlighting the role that small changes within the MTC have played in altering host preferences and geographical restriction.

In addition, we found that in Mali, *M. africanum*-associated and *M. tuberculosis*-associated strains evolved antibiotic resistance through similar mutations. This adds support to the hypothesis that these strains are undergoing similar evolutionary pressures in Mali. Furthermore, we found a *gidB* polymorphism not previously described which may account for much of the streptomycin resistance in Mali.

One hypothesis for the geographic restriction of lineages 5 and 6 is that they are less fit, either for transmission or in-host virulence. Several papers have shown no difference in transmission rates between *M. tuberculosis*-associated strains and *M. africanum*-associated strains [136,140,212,213]. Our Mali collection revealed three pairs of lineage 6 strains separated by 10 or fewer SNPs when aligned to *M. tuberculosis* H37Rv suggesting recent transmission of strains between patients [185]. One transmission event was between two HIV negative patients, one was between two HIV positive patients and one between an HIV positive patient and an HIV negative patient, suggesting that a compromised immune system is not required for a transmission event. However, all but one patient in our collection was native to Mali, so the ability to transmit may be specific to ethnic backgrounds prevalent in Mali. Furthermore, our ANI data demonstrated that there is comparable diversity in lineages 4 and 6, suggesting that

lineage 6 has not obviously undergone selective pressures that were not also experienced by other lineages in this region. Thus, our genomic study did not find evidence of reduced transmission or diversity as might be expected if lineages 5 and 6 were less fit to cause disease and transmit among people within this region of the world.

Loss of fitness could also result either in an inability to succeed as well as other lineages outside of West Africa, or it could lead to an inability to grow in hosts not of West African ethnicity. It has been hypothesized that *M. africanum* is less virulent within both humans, mice and guinea pigs than is *M. tuberculosis* [138,140-142]. As a result, we looked for polymorphisms in known virulence genes. Our results showed that although lineage 6 contains lineage-specific mutations in genes previously shown to be required for growth *in vitro*, in macrophages, and in mice, it is not enriched for mutations in these categories compared to lineages 1, 2 and 3. This suggests that this lineage is not geographically restricted because of an overall numerical loss of virulence or growth-associated genes.

Although there were no enrichments for mutations in these pathways, individual mutations can still greatly affect disease outcome, and analysis of our lineage-specific mutations identified several potential mechanisms that could lead to changes in how lineage 5 and 6 proliferate and cause disease. Since our assemblies were of very high quality, we were able to observe changes in genes that previous studies could not, thus providing a prioritized list of these genes for investigating lineage 5 and 6 characteristics. One of these mechanisms involves alterations of proteins exposed to the immune system, especially those involved in ESX secretion, which may explain the altered response to ESAT-6 in patients infected by lineage 6 [136]. Mycobacteria have five ESX secretion systems, also known as type VII secretion systems, which secrete small proteins across the bacterial cell envelope [214,215]. These secretion

systems are important to mycobacterial virulence. For example, ESX-1 secretion is lost as part of RD1 in *M. bovis* BCG vaccine strains, resulting in loss of ESAT-6 and CFP-10 secretion, and thus attenuation of the bacterium [216,217]. Thus, ESX secretion systems and their substrates play a crucial role in the interaction with the host immune system. We identified lineage-specific mutations in ESX genes in every lineage, indicating that each lineage may interact uniquely with the host immune system. Together, these mutations could lead to alterations in the pathogen-host immune interaction, resulting in a requirement for the West African immune system.

Genes involved in bacterial growth, including genes in the MCE operons and L,D-transpeptidases, also contained lineage 5 and 6-specific mutations. The MCE operons play an important role in the virulence of M. tuberculosis, particularly in mycobacterial growth in macrophages [188,205]. Lineages 5 and 6 contained mutations affecting the function of proteins in both operons 1 and 3, suggesting potential defects in growth within the host. Similarly, L,D-transpeptidases are critical to the structure of mycobacterial peptidoglycan, and loss of these enzymes affects bacterial structure and growth [204]. Two of the five L,D-transpeptidases in *M. tuberculosis* contained lineage 6-specific mutations affecting protein function, and one of these was also mutated in lineage 5, providing a possible explanation for the reported changes in cell size and doubling time in *M. africanum* GM041182 compared to *M. tuberculosis* H37Rv [138].

In addition, mutations in molybdenum and cobalamin metabolism and cAMP signaling could result in an altered niche for these lineages. The MTC has an expanded set of genes involved in molybdenum cofactor biosynthesis, which is involved in anaerobic respiration and the stress response, suggesting an important role for this cofactor in the evolution and biology of mycobacteria [207,218]. Likewise, vitamin B12 synthesis, which is found only in select mycobacteria, has specifically evolved in mycobacteria, which also have a transport system for

this vitamin, suggesting a crucial role for vitamin B12 in M. tuberculosis infection [208,219,220]. Thus, the presence of pseudogenes and lineage-specific non-synonymous SNPs that affect protein function in the synthesis of these cofactors in lineages 5 and 6 suggest that these strains are adapted for a different niche than lineages 1-4. In line with this, M. tuberculosis encodes 17 adenylate cyclases, emphasizing the importance of this molecule to the bacteria [221], but both lineages 5 and 6 contained lineage-specific mutations that affected the protein function of several adenylate cyclases, suggesting altered cAMP signaling in these strains.

In addition, we saw high variability in PE, PPE and PE-PGRS genes, including changes in gene content. These repetitive regions are difficult to sequence and so are often ignored, but may play a crucial role in antigenicity and the host-pathogen interaction [86,222]. However, using our high quality assemblies and alignments, we were able to identify lineage-specific mutations in these genes, as well as changes in gene content. These mutations highlight the possibility of a critical role in host-pathogen interactions and emphasize the need for a more detailed analysis of these regions. Furthermore, there were also a number of mutated hypothetical proteins and proteins of unknown function, all of which may play a critical as yet undiscovered role. Both repetitive genes and hypothetical proteins require further follow-up to elucidate their role in the differences between lineages.

A third hypothesis for the geographical restriction of lineages 5 and 6 is the presence of an unknown non-human reservoir. Indeed, *M. africanum* has been found in animals, including monkeys, cows, pigs and hyrax [223-228]. Unfortunately, given genomic data from human clinical isolates alone, we cannot address this hypothesis directly. However, given the similar level of diversity between lineage 4 and 6 in Mali and the evidence of person-to-person

transmission, even if lineage 6 has an animal reservoir, it is also well adapted to spread in humans living in this geographic setting.

One weakness of our study was that we were limited in our sample size for lineage 5 and *M. bovis* strains. Our collection was not representative of *M. bovis* genomic diversity, as three of the four *M. bovis* strains in our analysis were *M. bovis BCG* strains, which are attenuated lab strains used for vaccines. However, we only used the *M. bovis* strains in our ANI and gene content analysis, and required that any observations be consistent with wild-type *M. bovis* sequence, AF2122/97. Our results corroborated all previous findings of regions of difference, providing support that using BCG strains did not grossly alter our conclusions. Another weakness was that since all lineage 5 and 6 isolates in our study, except *M. africanum* GM041182, came from Mali, some of our observations may be specific to Mali. In fact, all of our patients were born in Mali, with the exception of one Central African patient infected with a lineage 4 strain. However, our lineage 6 isolates were genetically diverse and represented multiple different spoligotypes. In addition, our Mali isolates from other lineages did not cluster separately from strains from South Africa. Thus, while some of our conclusions may not apply outside of Mali, our collection reflected substantial diversity and did not originate from a clonal outbreak.

This collection provides valuable insights into the distinguishing genomic features of *M. africanum.* Here, we have shown that lineage 6 in Mali appears able to spread through person-to-person transmission and has diversified as well as lineage 4. Furthermore, we did not identify an increased rate of mutation in virulence-associated genes, also partially ruling out the hypothesis of decreased virulence and loss of fitness. In addition, we have identified several potential mechanisms for the geographical restriction of lineages 5 and 6, which provide a guide

to future studies focusing on the effects of specific genes. Future work can use these observations to inform experiments on mycobacterial pathogenicity and virulence, particularly with regard to this unique member of the MTC.

5.6 Figures

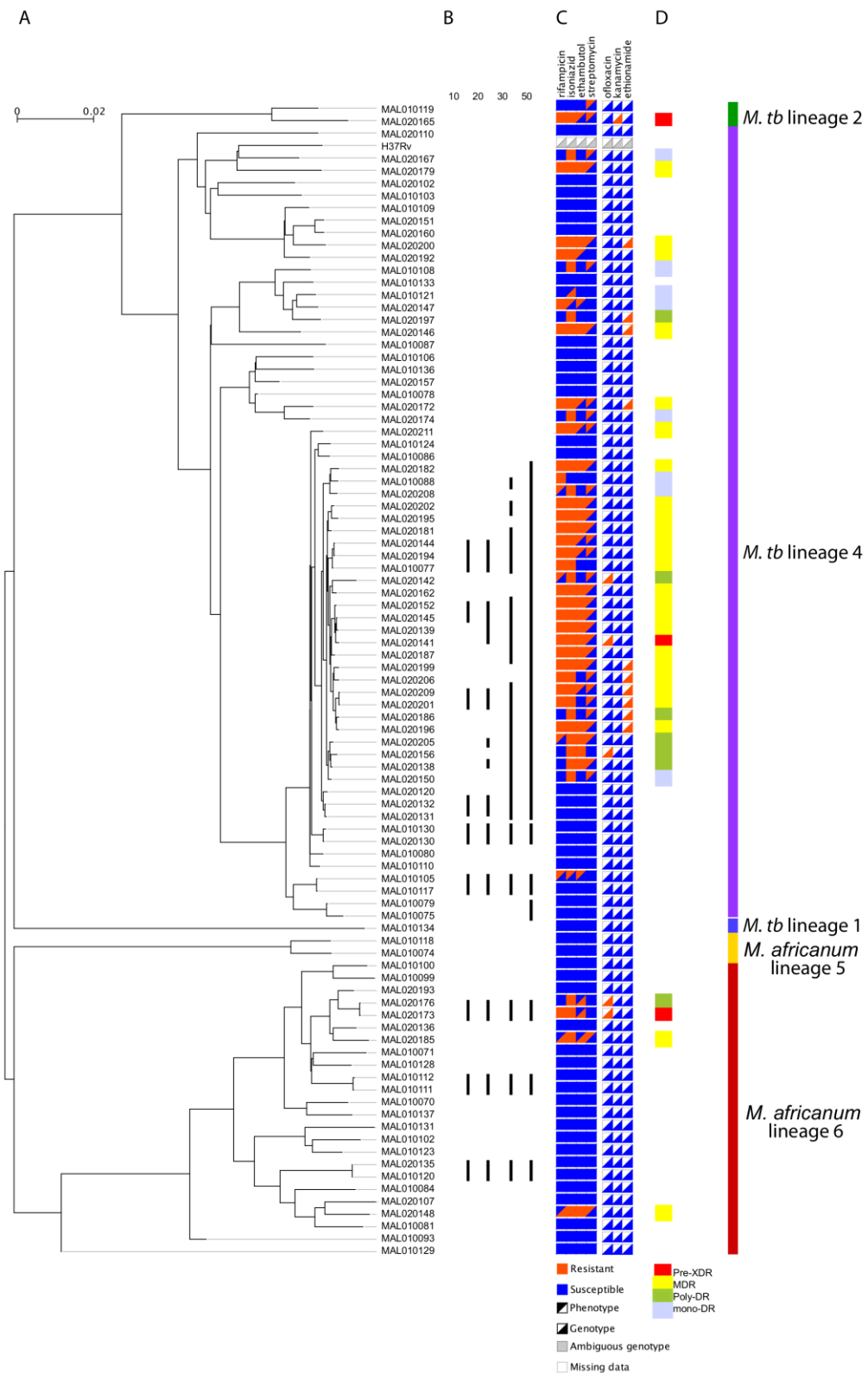

A
B
C
D
0
0.02
10
20
30
50
rifampicin
isoniazid
ethambutol
streptomycin
ofloxacin
kanamycin
ethionamide
MAL010119
MAL020165
MAL020110
H37Rv
MAL020167
MAL020179
MAL020102
MAL010103
MAL010109
MAL020151
MAL020160
MAL020200
MAL020192
MAL010108
MAL010133
MAL010121
MAL020147
MAL020197
MAL020146
MAL010087
MAL010106
MAL010136
MAL020157
MAL010078
MAL020172
MAL020174
MAL020211
MAL010124
MAL010086
MAL020182
MAL010088
MAL020208
MAL020202
MAL020195
MAL020181
MAL020144
MAL020194
MAL010077
MAL020142
MAL020162
MAL020152
MAL020145
MAL020139
MAL020141
MAL020187
MAL020199
MAL020206
MAL020209
MAL020201
MAL020186
MAL020196
MAL020205
MAL020156
MAL020138
MAL020150
MAL020120
MAL020132
MAL020131
MAL010130
MAL020130
MAL010080
MAL010110
MAL010105
MAL010117
MAL010079
MAL010075
MAL010134
MAL010118
MAL010074
MAL010100
MAL010099
MAL020193
MAL020176
MAL020173
MAL020136
MAL020185
MAL010071
MAL010128
MAL010112
MAL010111
MAL010070
MAL010137
MAL010131
MAL010102
MAL010123
MAL020135
MAL010120
MAL010084
MAL020107
MAL020148
MAL010081
MAL010093
MAL010129
M. tb lineage 2
M. tb lineage 4
M. tb lineage 1
M. africanum lineage 5
M. africanum lineage 6
Resistant
Susceptible
Phenotype
Genotype
Ambiguous genotype
Missing data
Pre-XDR
MDR
Poly-DR
mono-DR

Figure 5-1. *M. africanum* and *M. tuberculosis* drug resistance is genetically similar.

A) SNP-based phylogenetic tree of newly sequenced strains from Mali, constructed using FastTree [180]. B) Groups differing by 10, 20, 30, or 50 SNPs are shown with light grey bars, as calculated in Cohen et al. [154]. C) Comparison of genotypic and phenotypic DST (drug susceptibility testing). Genotypic drug resistance was calculated using HAIN genetic markers [190,191]. D) Drug resistance category (mono-DR, poly-DR, MDR, or pre-XDR) based on genotype.

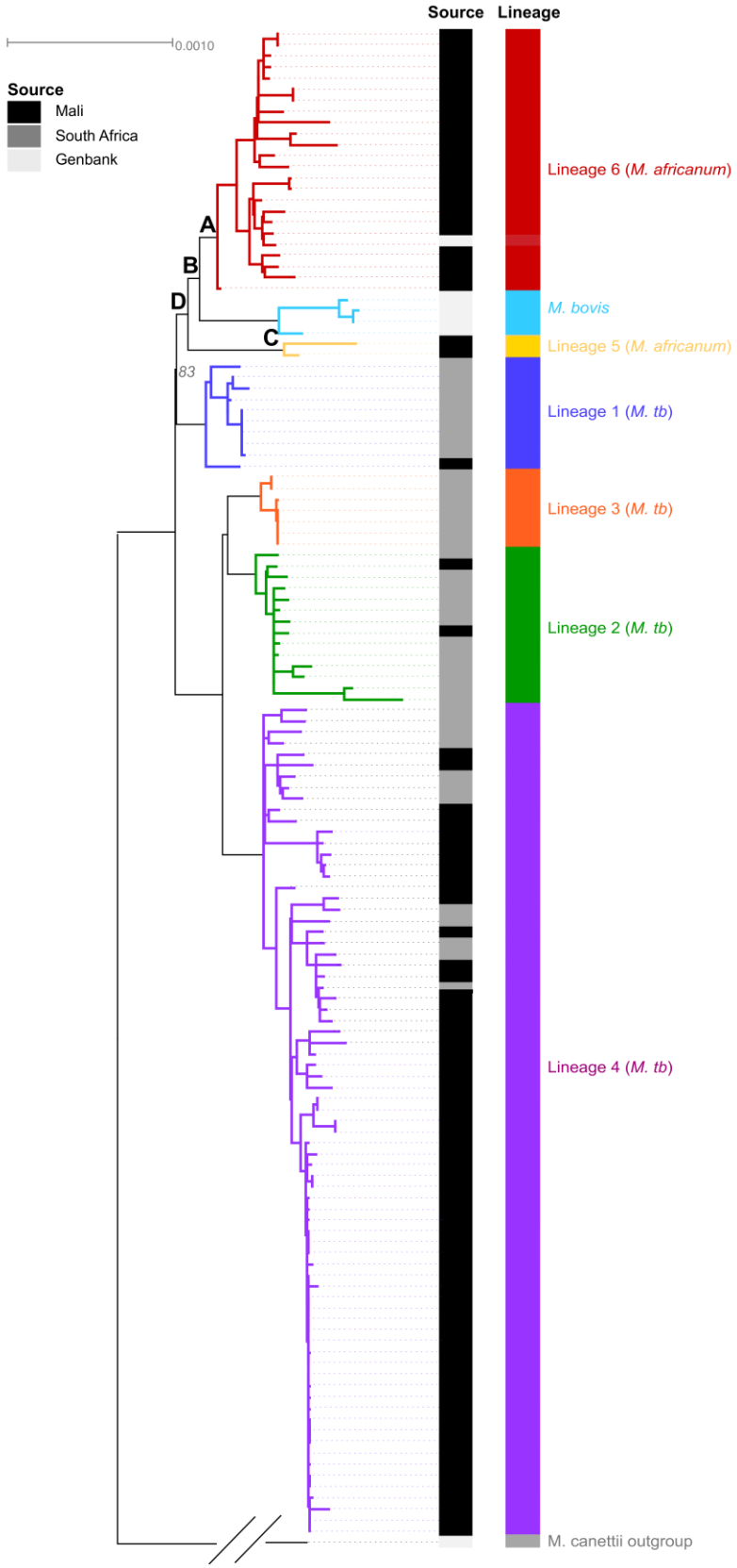

Source
Lineage
0.0010
Source
Mali
South Africa
Genbank
A
B
D
C
83
Lineage 6 (M. africanum)
M. bovis
Lineage 5 (M. africanum)
Lineage 1 (M. tb)
Lineage 3 (M. tb)
Lineage 2 (M. tb)
Lineage 4 (M. tb)
M. canettii outgroup

Figure 5-2. Phylogenetic tree of assembly collection.

Nodes, lineages, and Mali strains are indicated. All key nodes separating the major lineages had bootstrap values of 100%, except for the node separating *M. tuberculosis* lineage 1 and *M. africanum* lineage 5, which had a bootstrap value of 83%.

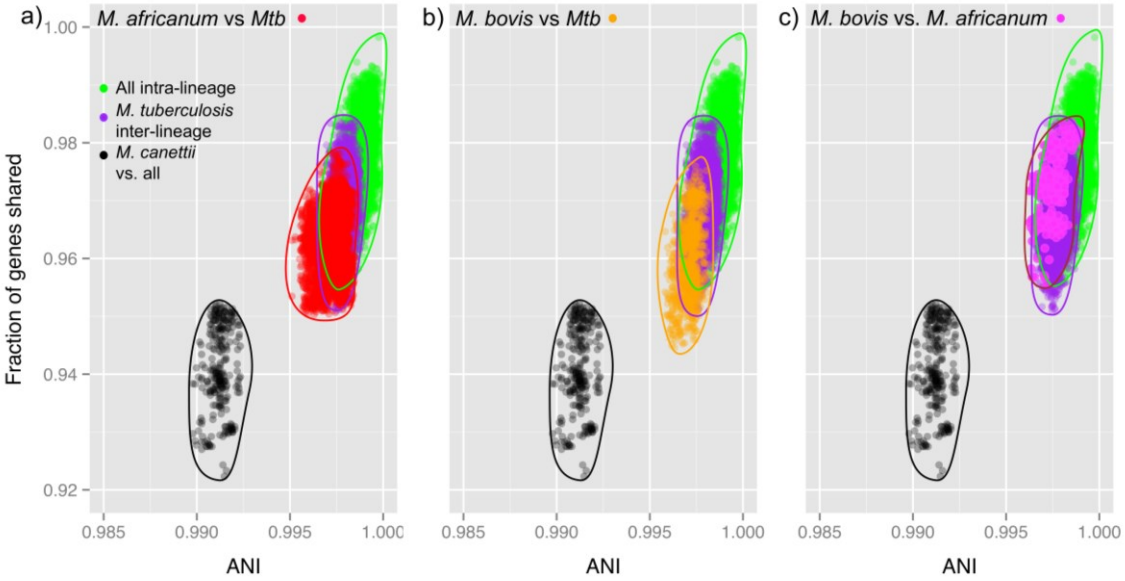


Figure 5-3. Average nucleotide identity (ANI) analysis indicates *M. africanum* and *M. tuberculosis* are not separate species.

A) ANI values when comparing *M. africanum* and *M. tuberculosis* do not cross the ANI species threshold of 94-95%. In fact, this comparison shows that the distribution of *M. africanum*/*M. tuberculosis* comparisons (red) overlaps that of inter-lineage *M. tuberculosis* comparisons (purple), indicating that *M. africanum* should be considered another lineage of *M. tuberculosis*. B) Similarly, ANI values when comparing *M. bovis* and *M. tuberculosis* also overlap with inter-lineage *M. tuberculosis,* and indicate that *M. bovis* should also be considered another lineage of *M. tuberculosis.* C) ANI values comparing *M. africanum* and *M. bovis* (pink) also overlap inter-lineage *M. tuberculosis* comparisons (green).

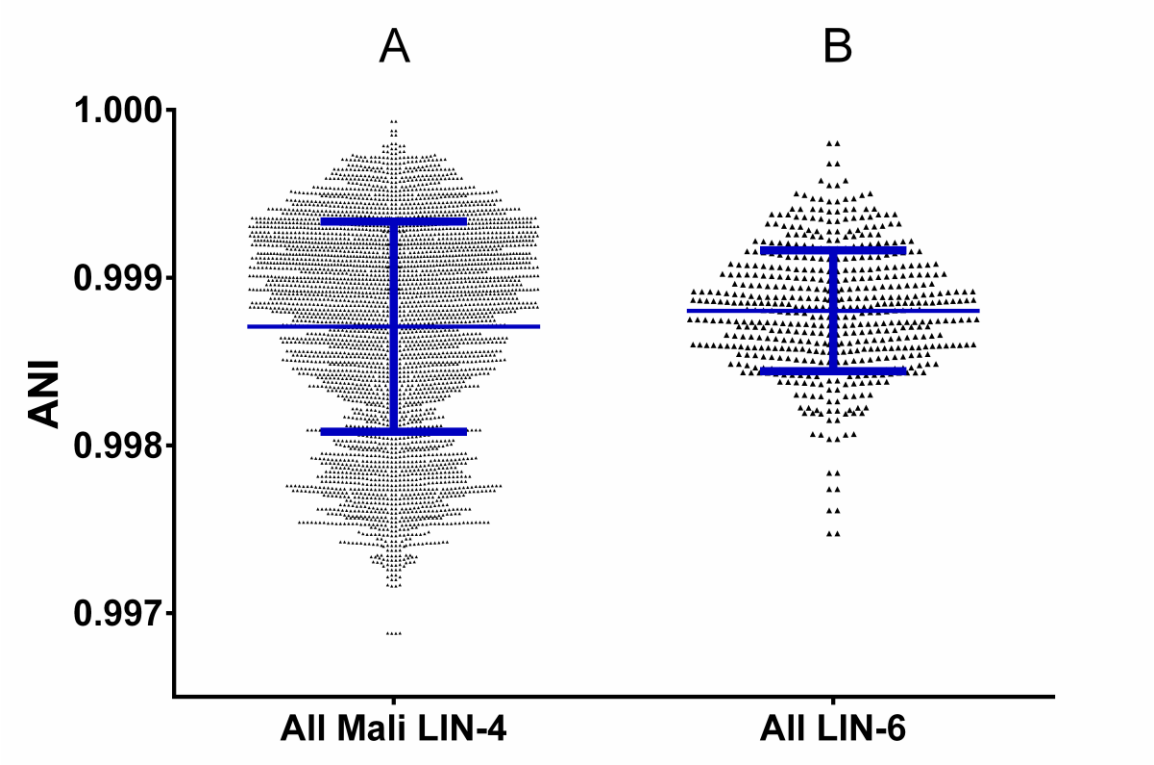


Figure 5-4. Diversity in Mali lineage 4 and lineage 6 strains diversity.

ANI values for comparisons (A) within all Mali lineage 4 isolates and (B) within all lineage 6 isolates. Blue lines indicate mean ± standard deviation. The means of these two groups was not significantly different.

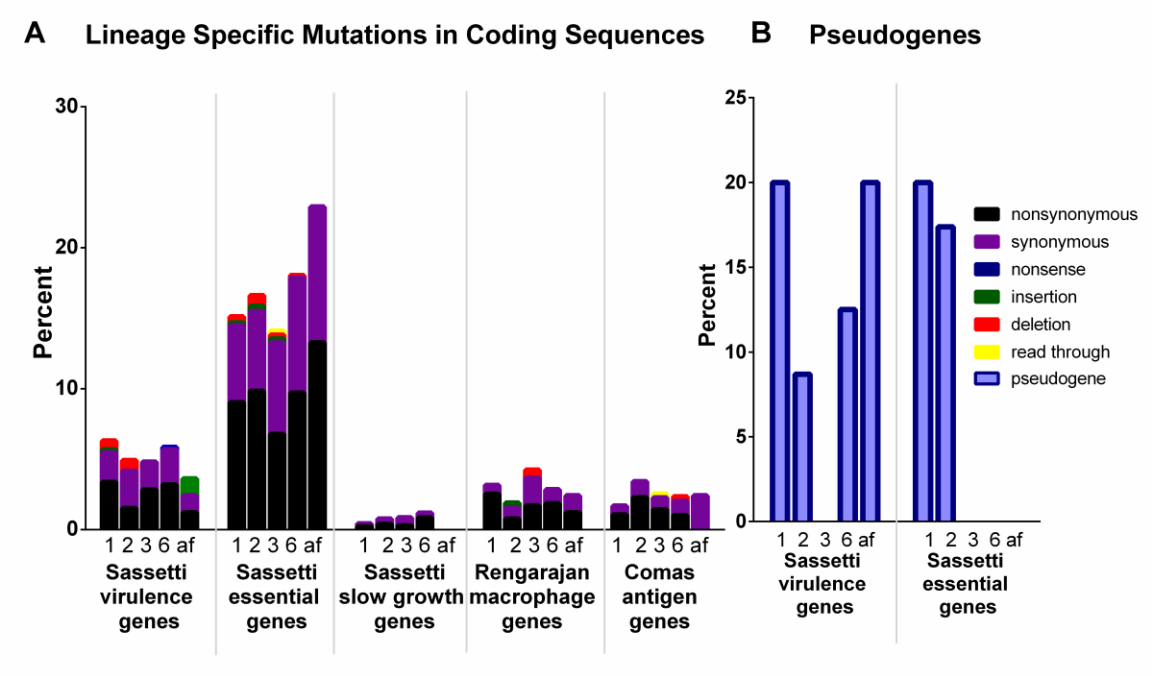


Figure 5-5. Percentage of lineage-specific mutations in virulence associated genes.

A) Percentage of lineage-specific mutations in coding sequences of the genes in each category. Sassetti virulence genes are genes that were identified in [187] as being required for virulence in mice. TraSH essential and slow growth genes were identified by Sassetti et al. under *in vitro* conditions [186]. Rengarajan macrophage genes were identified by Rengarajan et al. as being required for growth in macrophages [188]. Comas antigen genes were genes identified as containing T cell epitopes [135]. The color of the bar indicates type of mutation. B) Percentage of lineage-specific pseudogenes falling into the above defined categories. TraSH virulence genes and Comas antigen genes had no pseudogenes in any lineage. Lineage is indicated by the number below each bar, while af indicates mutations found in both lineage 5 and 6 (both *M. africanum* lineages).

5.7 Tables

	True positives	False negatives	False Positives	True Negatives	Sensitivity	Specificity	F	Total strains
rifampicin	28	4	2	58	87.50%	96.67%	90.32%	92
isoniazid	40	2	1	49	95.24%	98%	96.39%	92
ethambutol	19	9	3	61	67.86%	95.31%	76%	92
streptomycin	0	35	0	57	0%	100%	0%	92

Table 5-1. Genotypic drug resistance analysis.

Table showing true positives, false positives, true negatives, false negatives, sensitivity, and specificity for the four drugs for which we have phenotype information for all strains.

A: root node of lineage 6

loss	annotation	RD
Rv0124	PE-PGRS family protein PE_PGRS2	RD701

B: root node of lineage 6 and *M. bovis*

loss	annotation	RD
Rv0222	enoyl-CoA hydratase EchA1	RD10
Rv1965	ABC transporter permease YrbE3B	RD7
Rv1966	MCE-family protein Mce3A	RD7
Rv1967	MCE-family protein Mce3B	RD7
Rv1968	MCE-family protein Mce3C	RD7
Rv1969	MCE-family protein Mce3D	RD7
Rv1970	MCE-family lipoprotein LprM	RD7
Rv1971	MCE-family protein Mce3F	RD7
Rv1972	MCE-associated membrane protein	RD7
Rv1973	MCE-associated membrane protein	RD7
Rv1974	membrane protein	RD7
Rv1975	hypothetical protein	RD7
Rv1976c	hypothetical protein	RD7
Rv3617	epoxide hydrolase EphA	RD8
Rv3618	monooxygenase	RD8
Rv3619c	ESAT-6 like protein EsxV	RD8
Rv3620c	ESAT-6 like protein EsxW	RD8
Rv3621c	PPE family protein PPE65	RD8
Rv3622c	PE family protein PE32	RD8

gained	annotation	RD
Orthogroup 850447630	Hypothetical protein (has 2 domains: 104 kDa microneme/rhoptry antigen & large tegument protein UL36)	
Orthogroup 850450572	PE-PGRS family protein	

C: root node of lineage 5

loss	annotation	RD
Rv1334	hydrolase	RD711
Rv1335	sulfur carrier protein CysO	RD711
Rv1523	methyltransferase	
Rv1978	hypothetical protein	RD713
Rv1979c	permease	RD713
Rv1993c	hypothetical protein	RD743
Rv1994c	ArsR family transcriptional regulator CmtR	RD743

Rv1995	hypothetical protein	RD743
Rv3514	PE-PGRS family protein PE_PGRS57	

D: root node of lineages 5 and 6 and *M. bovis*

loss	**annotation**	**RD**
Rv2073c	oxidoreductase	RD9
Rv2074	pyridoxamine 5'-phosphate oxidase	RD9
Rv2084	hypothetical protein	
gained	**annotation**	**RD**
Orthogroup 850451604	PPE family protein	

Table 5-2: Orthologs identified in gene content analysis as lost or gained at nodes A-D.

For some RDs, the first and/or last gene in the region was not identified in our analysis because enough of the gene remained to align to H37Rv, and thus was not considered absent. Nodes A-D were identified in Fig. 2.

lineage	shared hits	our analysis only	Bentley analysis only
LIN-1	NA	5	NA
LIN-2	NA	23	NA
LIN-3	NA	5	NA
LIN-4	0	0	3
LIN-6	5	3	27
Maf	5	0	8

Table 5-3. Table comparing the pseudogenes identified in our study to those identified by Bentley et al. [148].

	Mutations		Pseudogenes	
Lineage	Average ± SD	Lineage Specific	Average ± SD	Lineage Specific
LIN-1	2527.2 ± 39.2	536	189 ± 21.5	5
LIN-2	1774.4 ± 64.9	308	183.5 ± 43.1	23
LIN-3	1740.0 ± 52.9	406	157.9 ± 46.5	5
LIN-4	1050.0 ± 143.9	1	99.5 ± 28.8	0
LIN-5	2540.5 ± 4.9	952	190.5 ± 0.7	43
LIN-6	2605.8 ± 82.0	681	201.5 ± 26.4	8
LIN-5 and LIN-6	2600.4 ± 80.5	90	200.5 ± 25.4	5
LIN-1, LIN2, LIN-3, and LIN-4	1355.4 ± 400.0	0	148.4 ± 55.4	0
LIN-1, LIN2, and LIN-3	1803.4 ± 161.9	NA	182.3 ± 42.5	NA

Table 5-4. Summary of the lineage-specific mutations and pseudogenes detected for each lineage.

Lineage 5 numbers are greyed out due to low sample size, causing low confidence in our results.

	Category:	ESX Secretion	L,D transpeptid-ase	mammalian cell entry (MCE)	molybdenum, riboflavin and cobalamin metabolism	adenylate cyclase	drug-resistance associated genes
Lineage 1	Indel	0	0	0	1	0	0
	Synonymous	3	0	1	2	1	0
	Non-synonymous	4	0	1	3	1	1
	Non-synonymous & affects protein function	1	0	1	2	0	1
Lineage 2	Indel	1	0	0	1	0	0
	Synonymous	2	0	0	2	0	0
	Non-synonymous	7	0	0	0	0	0
	Non-synonymous & affects protein function	3	0	0	0	0	0
Lineage 3	Indel	0	0	0	0	0	0
	Synonymous	2	0	2	1	2	1
	Non-synonymous	3	0	1	0	0	1
	Non-synonymous & affects protein function	0	0	1	0	0	0
Lineage 5	Indel	3	0	1	1	0	0
	Synonymous	10	0	0	7	2	1
	Non-synonymous	16	1	6	8	2	2
	Non-synonymous & affects protein function	7	1	4	4	2	2
Lineage 6	Indel	2	0	0*	0	1	0
	Synonymous	1	0	1	7	0	0
	Non-synonymous	7	2	3	7	2	3
	Non-synonymous & affects protein	1	1	1	3	0	0

	function						
Lineage 5 & 6	Indel	0	0	0	2	0	0
	Synonymous	2	0	0	0	0	0
	Non-synonymous	0	0	1	0	0	0
	Non-synonymous & affects protein function	0	0	1	0	0	0

Table 5-5. Summary of lineage-specific mutations of highlighted in 5.4.6.

Each heading refers to the results section these mutations are discussed.

Supplementary files

S1 Figure. Mutations in drug-resistance associated genes. Plots showing details of known drug resistance mutations present for each drug. Light blue or red horizontal shaded bars indicate phenotypic sensitivity or phenotypic resistance for the strain of interest. The corresponding vivid color in a particular box indicates the presence of the resistance mutation represented by that column. A) rifampicin B) isoniazid C) ethambutol D) streptomycin

S1 Table. Samples used in our study. A) List of Mali samples used in our study, with patient information. B) List of all 137 strains used for our assembly-based analyses, including 91 newly sequenced strains from Mali, 40 strains from the K-RITH collection from South Africa [154], and six strains from Genbank. C) Sequence Read Archive identifiers for each of the 161 additional strains from China used in our SNP analysis [156].

S2 Table. Drug resistance analysis. A) Drug resistance mutations analyzed. B) Table showing true positives, false positives, true negatives, false negatives, sensitivity, and specificity for the four drugs for which we have phenotype information for all strains.

S3 Table. Lineage-specific mutations. All lineage-specific mutations (A) in coding sequences and (B) in intergenic regions. Maf indicates mutations shared between lineage 5 and 6 but not found in lineages 1-4. No mutations were shared between lineages 1-4 but not lineages 5-6.

S4 Table. Lineage-specific pseudogenes. Maf indicates mutations shared between lineage 5 and 6 but not found in lineages 1-4. No mutations were shared between lineages 1-4 but not lineages 5-6.

S5 Table. Comparison of pseudogenes to previous analysis. A) Comparison of pseudogenes identified differently by our study to those identified by Bentley et al [148]. This table compares lineage 4, lineage 6 or *M. africanum*-specific pseudogenes identified in our study to pseudogenes identified by Bentley et al. as belonging to lineage 4, lineage 6, lineage 6 and animal strains or lineage 5, -6 and animal strains. A "0" indicates that the gene is not a pseudogene in that strain, while "1" indicates that it is, and "2" indicates an ambiguous call. Genes with a light blue background were identified in this study and not by Bentley et al., while genes with a light green background were identified by Bentley et al., but not by this study, and genes with a purple background were identified by both studies. B) Table summarizing the differences between our study and Bentley et al. [148].

Appendix

Legends for Supplemental Figures

Supplemental Figure 3-S1. Changes in gut microbiota gene expression with mycobacterial infection.

Heatmaps of log_2 fold-change and adjusted P-value compared to Day -3 for A) Balb/c, B) Black/6, C) $RAG^{-/-}$ and D) $MyD88^{-/-}$. All KEGG orthologs significant in at least one time-point are included. E) Shows the overlap between these groups.

Supplemental Figure 5-S1. Mutations in drug-resistance associated genes.

Plots showing details of known drug resistance mutations present for each drug. Light blue or red horizontal shaded bars indicate phenotypic sensitivity or phenotypic resistance for the strain of interest. The corresponding vivid color in a particular box indicates the presence of the resistance mutation represented by that column. A) rifampicin B) isoniazid C) ethambutol D) streptomycin

Legends for Supplemental Tables

Supplemental Table 2-S1. Differentially abundant OTUs in pre-infected samples and post-infected samples.

A q-value cutoff of $q<0.01$ was used.

Supplemental Table 2-S2. Differentially abundant OTUs in uninfected and infected samples.

A q-value cutoff of q<0.01 was used.

Supplemental Table 3-S1. OTUs and KEGG orthologs associated with *M. tuberculosis* colony forming unit (CFU) counts.

Analysis indicates which form of sequencing was used to calculate the relative abundance of the feature and organ indicates which organ the feature was associated with. A cutoff of Q<0.05 was used for this table.

Supplemental Table 3-S2. OTUs and KEGG orthologs associated cytokine levels.

Analysis indicates which form of sequencing was used to calculate the relative abundance of the feature, cytokine and organ indicates which cytokine and organ the feature was associated with. A cutoff of Q<0.05 was used for this table.

Supplemental Table 3-S3. OTUs identified through 16S rDNA sequencing significantly associated within infecting organism.

Y=yes, N=no, NA=not applicable. A cutoff of Q<0.05 was used for this table.

Supplemental Table 3-S4. OTUs identified through whole genome sequencing significantly associated within infecting organism.

Y=yes, N=no, NA=not applicable. A cutoff of Q<0.05 was used for this table.

Supplemental Table 3-S5. KEGG orthologs identified through 16S rDNA sequencing significantly associated within infecting organism.

Y=yes, N=no, NA=not applicable. A cutoff of Q<0.05 was used for this table. Outliers removed from these results.

Supplemental Table 3-S6. KEGG orthologs identified through whole genome sequencing significantly associated within infecting organism.

Y=yes, N=no, NA=not applicable. A cutoff of Q<0.05 was used for this table.

Supplemental Table 3-S7. KEGG orthologs identified through RNA sequencing with significant change in expression compared to pre-infection in at least one post-infection timepoint.

A cutoff of adjusted P <0.05 in at least one timepoint was used for this table.

Supplemental Table 5-S1. List of Mali samples used in our study, with patient information.

Supplemental Table 5-S2. List of all 137 strains used for our assembly-based analyses.

Includes 91 newly sequenced strains from Mali, 40 strains from the K-RITH collection [154], and six strains from Genbank.

Supplemental Table 5-S3. Sequence Read Archive identifiers for each of the 161 additional strains used in our SNP analysis.

Supplemental Table 5-S4. Drug resistance mutations analyzed.

Supplemental Table 5-S5. All lineage-specific mutations in coding sequences.

Supplemental Table 5-S6. All lineage-specific mutations in intergenic regions.

Supplemental Table 5-S7. Lineage specific pseudogenes.

Supplemental Table 5-S8. Comparison of pseudogenes identified differently by our study to previous analysis.

This table compares lineage 4, lineage 6 or M. africanum specific pseudogenes identified in our study to pseudogenes identified by Bentley et al. [148] as belonging to lineage 4, lineage 6, lineage 6 and animal strains or lineage 5, -6 and animal strains. A “0” indicates that the gene is

not a pseudogene in that strain, while "1" indicates that it is, and "2" indicates an ambiguous call. Genes with a light blue background were identified in this study and not by Bentley et al., while genes with a light green background were identified by Bentley et al., but not by this study, and genes with a purple background were identified by both studies.

Supplemental Table 5-S9. Lineage-specific mutations of highlighted in 5.4.6.

Each heading refers to the results section these mutations are discussed. For full list of lineage-specific mutations, see Supplemental Table 5-S5.

References

1. World Health Organization (2014) Global Tuberculosis Report. http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809_eng.pdf?ua=1
2. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, et al. (2013) Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45: 1176-1182.
3. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, et al. (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514: 494-497.
4. Zimmerman MR (1979) Pulmonary and osseous tuberculosis in an Egyptian mummy. Bull N Y Acad Med 55: 604-608.
5. Migliori GB, De Iaco G, Besozzi G, Centis R, Cirillo DM (2007) First tuberculosis cases in Italy resistant to all tested drugs. Euro Surveill 12: E070517 070511.

6. Parida SK, Axelsson-Robertson R, Rao MV, Singh N, Master I, et al. (2014) Totally drug-resistant tuberculosis and adjunct therapies. J Intern Med.

7. Dheda K, Ruhwald M, Theron G, Peter J, Yam WC (2013) Point-of-care diagnosis of tuberculosis: past, present and future. Respirology 18: 217-232.

8. Helb D, Jones M, Story E, Boehme C, Wallace E, et al. (2010) Rapid detection of Mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology. J Clin Microbiol 48: 229-237.

9. Ekblom R, Wolf JB (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7: 1026-1042.

10. Ohashi H, Hasegawa M, Wakimoto K, Miyamoto-Sato E (2015) Next-Generation Technologies for Multiomics Approaches Including Interactome Sequencing. Biomed Res Int 2015: 104209.

11. Nuermberger E (2008) Using animal models to develop new treatments for tuberculosis. Semin Respir Crit Care Med 29: 542-551.

12. Flynn JL, Chan J (2001) Tuberculosis: latency and reactivation. Infect Immun 69: 4195-4201.

13. Hooper LV, Littman DR, Macpherson AJ (2012) Interactions between the microbiota and the immune system. Science 336: 1268-1273.

14. Jarchum I, Pamer EG (2011) Regulation of innate and adaptive immunity by the commensal microbiota. Curr Opin Immunol 23: 353-360.

15. Maynard CL, Elson CO, Hatton RD, Weaver CT (2012) Reciprocal interactions of the intestinal microbiota and immune system. Nature 489: 231-241.

16. Nishio J, Honda K (2012) Immunoregulation by the gut microbiota. Cell Mol Life Sci 69: 3635-3650.

17. Penders J, Stobberingh EE, van den Brandt PA, Thijs C (2007) The role of the intestinal microbiota in the development of atopic disorders. Allergy 62: 1223-1236.

18. Qin J, Li Y, Cai Z, Li S, Zhu J, et al. (2012) A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature 490: 55-60.

19. Yeoh N, Burton JP, Suppiah P, Reid G, Stebbings S (2013) The role of the microbiome in rheumatic diseases. Curr Rheumatol Rep 15: 314.

20. Kuss SK, Best GT, Etheredge CA, Pruijssers AJ, Frierson JM, et al. (2011) Intestinal microbiota promote enteric virus replication and systemic pathogenesis. Science 334: 249-252.

21. Kamada N, Kim YG, Sham HP, Vallance BA, Puente JL, et al. (2012) Regulated virulence controls the ability of a pathogen to compete with the gut microbiota. Science 336: 1325-1329.

22. Sekirov I, Finlay BB (2009) The role of the intestinal microbiota in enteric infection. J Physiol 587: 4159-4167.

23. Wilks J, Golovkina T (2012) Influence of microbiota on viral infections. PLoS Pathog 8: e1002681.

24. Dubourg G, Lagier JC, Armougom F, Robert C, Hamad I, et al. (2013) The gut microbiota of a patient with resistant tuberculosis is more comprehensively studied by culturomics than by metagenomics. Eur J Clin Microbiol Infect Dis 32: 637-645.

25. Cui Z, Zhou Y, Li H, Zhang Y, Zhang S, et al. (2012) Complex sputum microbial composition in patients with pulmonary tuberculosis. BMC Microbiol 12: 276.

26. Cheung MK, Lam WY, Fung WY, Law PT, Au CH, et al. (2013) Sputum microbiota in tuberculosis as revealed by 16S rRNA pyrosequencing. PLoS One 8: e54574.

27. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods 5: 235-237.

28. Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One 6: e27310.

29. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.

30. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188-7196.

31. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73: 5261-5267.

32. Jari Oksanen FGB, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Helene Wagner vegan: Community Ecology Package.

33. White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol 5: e1000352.

34. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335-336.

35. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431-432.

36. Gill N, Wlodarska M, Finlay BB (2010) The future of mucosal immunology: studying an integrated system-wide organ. Nat Immunol 11: 558-560.

37. Chackerian AA, Alt JM, Perera TV, Dascher CC, Behar SM (2002) Dissemination of Mycobacterium tuberculosis is influenced by host factors and precedes the initiation of T-cell immunity. Infect Immun 70: 4501-4509.

38. Atarashi K, Tanoue T, Shima T, Imaoka A, Kuwahara T, et al. (2011) Induction of colonic regulatory T cells by indigenous Clostridium species. Science 331: 337-341.

39. Costa MC, Arroyo LG, Allen-Vercoe E, Stampfli HR, Kim PT, et al. (2012) Comparison of the fecal microbiota of healthy horses and horses with colitis by high throughput sequencing of the V3-V5 region of the 16S rRNA gene. PLoS One 7: e41484.

40. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, et al. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A 104: 13780-13785.

41. Russell SL, Gold MJ, Hartmann M, Willing BP, Thorson L, et al. (2012) Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma. EMBO Rep 13: 440-447.

42. Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL (2005) An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. Cell 122: 107-118.

43. Huda MN, Lewis Z, Kalanetra KM, Rashid M, Ahmad SM, et al. (2014) Stool microbiota and vaccine responses of infants. Pediatrics 134: e362-372.

44. Botero LE, Delgado-Serrano L, Cepeda ML, Bustos JR, Anzola JM, et al. (2014) Respiratory tract clinical sample selection for microbiota analysis in patients with pulmonary tuberculosis. Microbiome 2: 29.

45. Wu J, Liu W, He L, Huang F, Chen J, et al. (2013) Sputum microbiota associated with new, recurrent and treatment failure tuberculosis. PLoS One 8: e83445.

46. Littman DR, Pamer EG (2011) Role of the commensal microbiota in normal and pathogenic host immune responses. Cell Host Microbe 10: 311-323.

47. Kamada N, Seo SU, Chen GY, Nunez G (2013) Role of the gut microbiota in immunity and inflammatory disease. Nat Rev Immunol 13: 321-335.

48. Kamada N, Chen GY, Inohara N, Nunez G (2013) Control of pathogens and pathobionts by the gut microbiota. Nat Immunol 14: 685-690.

49. Chu H, Mazmanian SK (2013) Innate immune recognition of the microbiota promotes host-microbial symbiosis. Nat Immunol 14: 668-675.

50. Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, et al. (2014) The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe 15: 382-392.

51. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, et al. (2014) Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A 111: E2329-2338.

52. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, et al. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31: 814-821.

53. Winglee K, Eloe-Fadrosh E, Gupta S, Guo H, Fraser C, et al. (2014) Aerosol Mycobacterium tuberculosis infection causes rapid loss of diversity in gut microbiota. PLoS One 9: e97048.

54. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, et al. (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 8: e1002358.

55. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, et al. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9: 811-814.

56. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, et al. (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 13: R79.

57. Kenneth Murphy PT, Mark Walport (2008) Janeway's Immunobiology (7th edition): Garland Science.

58. Shiloh MU, Champion PA (2010) To catch a killer. What can mycobacterial models teach us about Mycobacterium tuberculosis pathogenesis? Curr Opin Microbiol 13: 86-92.

59. Henkle E, Winthrop KL (2015) Nontuberculous Mycobacteria Infections in Immunosuppressed Hosts. Clin Chest Med 36: 91-99.

60. Ivanov, II, Atarashi K, Manel N, Brodie EL, Shima T, et al. (2009) Induction of intestinal Th17 cells by segmented filamentous bacteria. Cell 139: 485-498.

61. Hildebrand F, Nguyen TL, Brinkman B, Yunta RG, Cauwe B, et al. (2013) Inflammation-associated enterotypes, host genotype, cage and inter-individual effects drive gut microbiota variation in common laboratory mice. Genome Biol 14: R4.

62. McCafferty J, Muhlbauer M, Gharaibeh RZ, Arthur JC, Perez-Chanona E, et al. (2013) Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. ISME J 7: 2116-2125.

63. Clavel T, Duck W, Charrier C, Wenning M, Elson C, et al. (2010) Enterorhabdus caecimuris sp. nov., a member of the family Coriobacteriaceae isolated from a mouse model of spontaneous colitis, and emended description of the genus Enterorhabdus Clavel et al. 2009. Int J Syst Evol Microbiol 60: 1527-1531.

64. Bode LM, Bunzel D, Huch M, Cho GS, Ruhland D, et al. (2013) In vivo and in vitro metabolism of trans-resveratrol by human gut microbiota. Am J Clin Nutr 97: 295-309.

65. Maruo T, Sakamoto M, Ito C, Toda T, Benno Y (2008) Adlercreutzia equolifaciens gen. nov., sp. nov., an equol-producing bacterium isolated from human faeces, and emended description of the genus Eggerthella. Int J Syst Evol Microbiol 58: 1221-1227.

66. World Health Organization (2015) Tuberculosis. http://www.who.int/mediacentre/factsheets/fs104/en/.

67. Buve A, Jespers V, Crucitti T, Fichorova RN (2014) The vaginal microbiota and susceptibility to HIV. AIDS 28: 2333-2344.

68. Vazquez-Castellanos JF, Serrano-Villar S, Latorre A, Artacho A, Ferrus ML, et al. (2014) Altered metabolism of gut microbiota contributes to chronic immune activation in HIV-infected individuals. Mucosal Immunol.

69. Voigt RM, Keshavarzian A, Losurdo J, Swanson G, Siewe B, et al. (2015) HIV-associated mucosal gene expression: region-specific alterations. AIDS 29: 537-546.

70. Mutlu EA, Keshavarzian A, Losurdo J, Swanson G, Siewe B, et al. (2014) A compositional look at the human gastrointestinal microbiome and immune activation parameters in HIV infected subjects. PLoS Pathog 10: e1003829.

71. Ellis CL, Ma ZM, Mann SK, Li CS, Wu J, et al. (2011) Molecular characterization of stool microbiota in HIV-infected subjects by panbacterial and order-level 16S ribosomal DNA (rDNA) quantification and correlations with immune activation. J Acquir Immune Defic Syndr 57: 363-370.

72. Gori A, Tincati C, Rizzardini G, Torti C, Quirino T, et al. (2008) Early impairment of gut function and gut flora supporting a role for alteration of gastrointestinal mucosa in human immunodeficiency virus pathogenesis. J Clin Microbiol 46: 757-758.

73. Lozupone CA, Li M, Campbell TB, Flores SC, Linderman D, et al. (2013) Alterations in the gut microbiota associated with HIV-1 infection. Cell Host Microbe 14: 329-339.

74. World Health Organization (2014) Tuberculosis. http://www.who.int/mediacentre/factsheets/fs104/en/

75. Pieroni M, Tipparaju SK, Lun S, Song Y, Sturm AW, et al. (2011) Pyrido[1,2-a]benzimidazole-based agents active against tuberculosis (TB), multidrug-resistant (MDR) TB and extensively drug-resistant (XDR) TB. ChemMedChem 6: 334-342.

76. Lew JM, Kapopoulou A, Jones LM, Cole ST (2011) TubercuList--10 years after. Tuberculosis (Edinb) 91: 1-7.

77. Cohen SP, Hachler H, Levy SB (1993) Genetic and functional analysis of the multiple antibiotic resistance (mar) locus in Escherichia coli. J Bacteriol 175: 1484-1492.

78. George AM, Levy SB (1983) Gene in the major cotransduction gap of the Escherichia coli K-12 linkage map required for the expression of chromosomal resistance to tetracycline and other antibiotics. J Bacteriol 155: 541-548.

79. Reverchon S, Nasser W, Robert-Baudouy J (1994) pecS: a locus controlling pectinase, cellulase and blue pigment production in Erwinia chrysanthemi. Mol Microbiol 11: 1127-1139.

80. Davis JR, Sello JK (2010) Regulation of genes in Streptomyces bacteria required for catabolism of lignin-derived aromatic compounds. Appl Microbiol Biotechnol 86: 921-929.

81. Providenti MA, Wyndham RC (2001) Identification and functional characterization of CbaR, a MarR-like modulator of the cbaABC-encoded chlorobenzoate catabolism pathway. Appl Environ Microbiol 67: 3530-3541.

82. Poole K, Srikumar R (2001) Multidrug efflux in Pseudomonas aeruginosa: components, mechanisms and clinical significance. Curr Top Med Chem 1: 59-71.

83. Mongkolsuk S, Praituan W, Loprasert S, Fuangthong M, Chamnongpol S (1998) Identification and characterization of a new organic hydroperoxide resistance (ohr) gene with a novel pattern of oxidative stress regulation from Xanthomonas campestris pv. phaseoli. J Bacteriol 180: 2636-2643.

84. Collins L, Franzblau SG (1997) Microplate alamar blue assay versus BACTEC 460 system for high-throughput screening of compounds against Mycobacterium tuberculosis and Mycobacterium avium. Antimicrob Agents Chemother 41: 1004-1009.

85. Larsen MH, Biermann K, Tandberg S, Hsu T, Jacobs WR, Jr. (2007) Genetic Manipulation of Mycobacterium tuberculosis. Curr Protoc Microbiol Chapter 10: Unit 10A 12.

86. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, et al. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393: 537-544.

87. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

88. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 11: 11 10 11-11 10 33.

89. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

90. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

91. Lee MH, Pascopella L, Jacobs WR, Jr., Hatfull GF (1991) Site-specific integration of mycobacteriophage L5: integration-proficient vectors for Mycobacterium smegmatis, Mycobacterium tuberculosis, and bacille Calmette-Guerin. Proc Natl Acad Sci U S A 88: 3111-3115.

92. G.F H, W.R.Jr. J (2000) Molecular Genetics of Mycobacteria. Washington D.C: ASM Press.

93. Hatfull GF, Jacobs WR, Jr (2000) RNA Preparation-Trizol; Hatfull GF, Jacobs WR, Jr, editors. Washington DC: American Society for Microbiology.

94. Manganelli R, Dubnau E, Tyagi S, Kramer FR, Smith I (1999) Differential expression of 10 sigma factor genes in Mycobacterium tuberculosis. Mol Microbiol 31: 715-724.

95. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, et al. (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43: D222-226.

96. Pei J, Tang M, Grishin NV (2008) PROMALS3D web server for accurate multiple protein sequence and structure alignments. Nucleic Acids Res 36: W30-34.

97. Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 4: 363-371.

98. Simossis VA, Heringa J (2005) PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res 33: W289-294.

99. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

100. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.

101. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4: 1073-1081.

102. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.

103. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140.

104. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515.

105. Lamichhane G, Zignol M, Blades NJ, Geiman DE, Dougherty A, et al. (2003) A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 100: 7213-7218.

106. Sulavik MC, Dazer M, Miller PF (1997) The Salmonella typhimurium mar locus: molecular and genetic analyses and assessment of its role in virulence. J Bacteriol 179: 1857-1866.

107. Alekshun MN, Levy SB, Mealy TR, Seaton BA, Head JF (2001) The crystal structure of MarR, a regulator of multiple antibiotic resistance, at 2.3 A resolution. Nat Struct Biol 8: 710-714.

108. Wilkinson SP, Grove A (2006) Ligand-responsive transcriptional regulation by members of the MarR family of winged helix proteins. Curr Issues Mol Biol 8: 51-62.

109. Perera IC, Grove A (2010) Molecular mechanisms of ligand-mediated attenuation of DNA binding by MarR family transcriptional regulators. J Mol Cell Biol 2: 243-254.

110. Alekshun MN, Levy SB (1999) The mar regulon: multiple resistance to antibiotics and other toxic chemicals. Trends Microbiol 7: 410-413.

111. Szumowski JD, Adams KN, Edelstein PH, Ramakrishnan L (2013) Antimicrobial efflux pumps and Mycobacterium tuberculosis drug tolerance: evolutionary considerations. Curr Top Microbiol Immunol 374: 81-108.

112. Dawkins MJ, Judah JD, Rees KR (1959) The mechanism of action of chlorpromazine. Reduced diphosphopyridine nucleotidecytochrome c reductase and coupled phosphorylation. Biochem J 73: 16-23.

113. Dawkins MJ, Judah JD, Rees KR (1959) The effect of chlorpromazine on the respiratory chain; cytochrome oxidase. Biochem J 72: 204-209.

114. Finkelstein A (1970) Weak-acid uncouplers of oxidative phosphorylation. Mechanism of action on thin lipid membranes. Biochim Biophys Acta 205: 1-6.

115. Ren Q, Chen K, Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res 35: D274-279.

116. Denkin S, Byrne S, Jie C, Zhang Y (2005) Gene expression profiling analysis of Mycobacterium tuberculosis genes in response to salicylate. Arch Microbiol 184: 152-157.

117. de Knegt GJ, Bruning O, ten Kate MT, de Jong M, van Belkum A, et al. (2013) Rifampicin-induced transcriptome response in rifampicin-resistant Mycobacterium tuberculosis. Tuberculosis (Edinb) 93: 96-101.

118. de Souza GA, Leversen NA, Malen H, Wiker HG (2011) Bacterial proteins with cleaved or uncleaved signal peptides of the general secretory pathway. J Proteomics 75: 502-510.

119. Malen H, Berven FS, Fladmark KE, Wiker HG (2007) Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv. Proteomics 7: 1702-1718.

120. Garbe TR (2004) Co-induction of methyltransferase Rv0560c by naphthoquinones and fibric acids suggests attenuation of isoprenoid quinone action in Mycobacterium tuberculosis. Can J Microbiol 50: 771-778.

121. Starck J, Kallenius G, Marklund BI, Andersson DI, Akerlund T (2004) Comparative proteome analysis of Mycobacterium tuberculosis grown under aerobic and anaerobic conditions. Microbiology 150: 3821-3829.

122. Bacon J, Dover LG, Hatch KA, Zhang Y, Gomes JM, et al. (2007) Lipid composition and transcriptional response of Mycobacterium tuberculosis grown under iron-limitation in continuous culture: identification of a novel wax ester. Microbiology 153: 1435-1444.

123. Gu S, Chen J, Dobos KM, Bradbury EM, Belisle JT, et al. (2003) Comprehensive proteomic profiling of the membrane constituents of a Mycobacterium tuberculosis strain. Mol Cell Proteomics 2: 1284-1296.

124. Johnston JM, Arcus VL, Morton CJ, Parker MW, Baker EN (2003) Crystal structure of a putative methyltransferase from Mycobacterium tuberculosis: misannotation of a genome clarified by protein structural analysis. J Bacteriol 185: 4057-4065.

125. Deb C, Daniel J, Sirakova TD, Abomoelak B, Dubey VS, et al. (2006) A novel lipase belonging to the hormone-sensitive lipase family induced under starvation to utilize stored triacylglycerol in Mycobacterium tuberculosis. J Biol Chem 281: 3866-3875.

126. Alekshun MN, Levy SB (1999) Alteration of the repressor activity of MarR, the negative regulator of the Escherichia coli marRAB locus, by multiple chemicals in vitro. J Bacteriol 181: 4669-4672.

127. Sun Z, Cheng SJ, Zhang H, Zhang Y (2001) Salicylate uniquely induces a 27-kDa protein in tubercle bacillus. FEMS Microbiol Lett 203: 211-216.

128. Schuessler DL, Parish T (2012) The promoter of Rv0560c is induced by salicylate and structurally-related compounds in Mycobacterium tuberculosis. PLoS One 7: e34471.

129. McDermott PF, White DG, Podglajen I, Alekshun MN, Levy SB (1998) Multidrug resistance following expression of the Escherichia coli marA gene in Mycobacterium smegmatis. J Bacteriol 180: 2995-2998.

130. Zhang H, Gao L, Zhang J, Li W, Yang M, et al. (2014) A novel marRAB operon contributes to the rifampicin resistance in Mycobacterium smegmatis. PLoS One 9: e106016.

131. Radhakrishnan A, Kumar N, Wright CC, Chou TH, Tringides ML, et al. (2014) Crystal structure of the transcriptional regulator Rv0678 of Mycobacterium tuberculosis. J Biol Chem 289: 16526-16540.

132. de Jong BC, Antonio M, Gagneux S (2010) Mycobacterium africanum--review of an important cause of human tuberculosis in West Africa. PLoS Negl Trop Dis 4: e744.

133. Castets M, Boisvert H, Grumbach F, Brunel M, Rist N (1968) [Tuberculosis bacilli of the African type: preliminary note]. Rev Tuberc Pneumol (Paris) 32: 179-184.

134. Mostowy S, Onipede A, Gagneux S, Niemann S, Kremer K, et al. (2004) Genomic analysis distinguishes Mycobacterium africanum. J Clin Microbiol 42: 3594-3599.

135. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, et al. (2010) Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. Nat Genet 42: 498-503.

136. de Jong BC, Hill PC, Brookes RH, Gagneux S, Jeffries DJ, et al. (2006) Mycobacterium africanum elicits an attenuated T cell response to early secreted antigenic target, 6 kDa, in patients with tuberculosis and their household contacts. J Infect Dis 193: 1279-1286.

137. Tientcheu LD, Sutherland JS, de Jong BC, Kampmann B, Jafali J, et al. (2014) Differences in T-cell responses between Mycobacterium tuberculosis and Mycobacterium africanum-infected patients. Eur J Immunol 44: 1387-1398.

138. Bold TD, Davis DC, Penberthy KK, Cox LM, Ernst JD, et al. (2012) Impaired fitness of Mycobacterium africanum despite secretion of ESAT-6. J Infect Dis 205: 984-990.

139. Gehre F, Otu J, DeRiemer K, de Sessions PF, Hibberd ML, et al. (2013) Deciphering the growth behaviour of Mycobacterium africanum. PLoS Negl Trop Dis 7: e2220.

140. de Jong BC, Hill PC, Aiken A, Awine T, Antonio M, et al. (2008) Progression to active tuberculosis, but not transmission, varies by Mycobacterium tuberculosis lineage in The Gambia. J Infect Dis 198: 1037-1043.

141. de Jong BC, Adetifa I, Walther B, Hill PC, Antonio M, et al. (2010) Differences between tuberculosis cases infected with Mycobacterium africanum, West African type 2, relative to Euro-American Mycobacterium tuberculosis: an update. FEMS Immunol Med Microbiol 58: 102-105.

142. Castets M, Sarrat H (1969) [Experimental study of the virulence of Mycobacterium africanum (preliminary note)]. Bull Soc Med Afr Noire Lang Fr 14: 693-696.

143. Meyer CG, Scarisbrick G, Niemann S, Browne EN, Chinbuah MA, et al. (2008) Pulmonary tuberculosis: virulence of Mycobacterium africanum and relevance in HIV co-infection. Tuberculosis (Edinb) 88: 482-489.

144. Coscolla M, Gagneux S (2014) Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol 26: 431-444.

145. Smith NH, Kremer K, Inwald J, Dale J, Driscoll JR, et al. (2006) Ecotypes of the Mycobacterium tuberculosis complex. J Theor Biol 239: 220-225.

146. Asante-Poku A, Yeboah-Manu D, Otchere ID, Aboagye SY, Stucki D, et al. (2015) Mycobacterium africanum Is Associated with Patient Ethnicity in Ghana. PLoS Negl Trop Dis 9: e3370.

147. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, et al. (2002) A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 99: 3684-3689.

148. Bentley SD, Comas I, Bryant JM, Walker D, Smith NH, et al. (2012) The genome of Mycobacterium africanum West African 2 reveals a lineage-specific locus and genome erosion common to the M. tuberculosis complex. PLoS Negl Trop Dis 6: e1552.

149. Traore B, Diarra B, Dembele BP, Somboro AM, Hammond AS, et al. (2012) Molecular strain typing of Mycobacterium tuberculosis complex in Bamako, Mali. Int J Tuberc Lung Dis 16: 911-916.

150. Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, et al. (2003) The complete genome sequence of Mycobacterium bovis. Proc Natl Acad Sci U S A 100: 7877-7882.

151. Orduna P, Cevallos MA, de Leon SP, Arvizu A, Hernandez-Gonzalez IL, et al. (2011) Genomic and proteomic analyses of Mycobacterium bovis BCG Mexico 1931 reveal a diverse immunogenic repertoire against tuberculosis infection. BMC Genomics 12: 493.

152. Brosch R, Gordon SV, Garnier T, Eiglmeier K, Frigui W, et al. (2007) Genome plasticity of BCG and impact on vaccine efficacy. Proc Natl Acad Sci U S A 104: 5596-5601.

153. Seki M, Honda I, Fujita I, Yano I, Yamamoto S, et al. (2009) Whole genome sequence analysis of Mycobacterium bovis bacillus Calmette-Guerin (BCG) Tokyo 172: a comparative study of BCG vaccine substrains. Vaccine 27: 1710-1716.

154. Cohen K, Abeel, T.; Manson McGuire,A; Desjardins,CA; Munsamy,V; Shea,TP; Walker, BJ.; Bantubani,N; Almeida,D; Alvarado,L; Chapman,S; Mvelase,NR; Duffy,EY;

FitzGerald,MG;Govender,P; Gujja,S; Hamilton,S; Howarth,C; Larimer,JD; Maharaj,K; Pearson,MD; Priest,ME;Zeng,Q; Padayatchi,N; Grosset,J;Young,SJ; Wortman,J; Mlisana K; O'Donnell,MR;Birren,BW;Bishai,WR;Pym,AS. (2015) Evolution of extensively drug-resistant tuberculosis over four decades revealed by whole genome sequencing of Mycobacterium tuberculosis from KwaZulu-Natal, South Africa. Lancet.

155. Supply P, Marceau M, Mangenot S, Roche D, Rouanet C, et al. (2013) Genomic analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of Mycobacterium tuberculosis. Nat Genet 45: 172-179.

156. Zhang H, Li D, Zhao L, Fleming J, Lin N, et al. (2013) Genome sequencing of 161 Mycobacterium tuberculosis isolates from China identifies genes and intergenic regions associated with drug resistance. Nat Genet 45: 1255-1260.

157. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9: e112963.

158. Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, et al. (2012) Finished bacterial genomes from shotgun sequence data. Genome Res 22: 2270-2277.

159. Fisher S, Barry A, Abreu J, Minie B, Nolan J, et al. (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol 12: R1.

160. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.

161. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics 11: 119.

162. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research 25: 955-964.

163. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, et al. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic acids research 35: 3100-3108.

164. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281-288.

165. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, et al. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29: 41-43.

166. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29-34.

167. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631-637.

168. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

169. Tian W, Arakaki AK, Skolnick J (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res 32: 6226-6239.

170. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785-786.

171. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580.

172. Griggs A, Wapinski, I., Wortman, J., Haas, B. (2014) SYNERGY2: Accurate and scalable ortholog identification. in preparation.

173. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics 23: i549-558.

174. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. Nature 449: 54-61.

175. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

176. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102: 2567-2572.

177. Palmer KL, Godfrey P, Griggs A, Kos VN, Zucker J, et al. (2012) Comparative genomics of enterococci: variation in Enterococcus faecalis, clade structure in E. faecium, and defining characteristics of E. gallinarum and E. casseliflavus. mBio 3: e00318-00311.

178. Wilgenbusch JC, Swofford D (2003) Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics Chapter 6: Unit 6 4.

179. Picard.

180. Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490.

181. Storey JD (2002) A direct approach to false discovery rates. J R Statist Soc B 64: 479-498.

182. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics 10: 296.

183. Koeck JL, Fabre M, Simon F, Daffe M, Garnotel E, et al. (2011) Clinical characteristics of the smooth tubercle bacilli 'Mycobacterium canettii' infection suggest the existence of an environmental reservoir. Clin Microbiol Infect 17: 1013-1019.

184. Smith NH, Gordon SV, de la Rua-Domenech R, Clifton-Hadley RS, Hewinson RG (2006) Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. Nat Rev Microbiol 4: 670-681.

185. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, et al. (2013) Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13: 137-146.

186. Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. Mol Microbiol 48: 77-84.

187. Sassetti CM, Rubin EJ (2003) Genetic requirements for mycobacterial survival during infection. Proc Natl Acad Sci U S A 100: 12989-12994.

188. Rengarajan J, Bloom BR, Rubin EJ (2005) Genome-wide requirements for Mycobacterium tuberculosis adaptation and survival in macrophages. Proc Natl Acad Sci U S A 102: 8327-8332.

189. Homolka S, Post E, Oberhauser B, George AG, Westman L, et al. (2008) High genetic diversity among Mycobacterium tuberculosis complex strains from Sierra Leone. BMC Microbiol 8: 103.

190. Hain Genotype MTBDRplus. http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/genotype-mtbdrplus.html

191. Hain Genotype MTBDRsl. http://www.hain-lifescience.de/en/products/microbiology/mycobacteria/genotype-mtbdrsl.html

192. Hillemann D, Rusch-Gerdes S, Richter E (2009) Feasibility of the GenoType MTBDRsl assay for fluoroquinolone, amikacin-capreomycin, and ethambutol resistance testing of Mycobacterium tuberculosis strains and clinical specimens. J Clin Microbiol 47: 1767-1772.

193. Ignatyeva O, Kontsevaya I, Kovalyov A, Balabanova Y, Nikolayevskyy V, et al. (2012) Detection of resistance to second-line antituberculosis drugs by use of the genotype MTBDRsl assay: a multicenter evaluation and feasibility study. J Clin Microbiol 50: 1593-1597.

194. Kiet VS, Lan NT, An DD, Dung NH, Hoa DV, et al. (2010) Evaluation of the MTBDRsl test for detection of second-line-drug resistance in Mycobacterium tuberculosis. J Clin Microbiol 48: 2934-2939.

195. Kontsevaya I, Ignatyeva O, Nikolayevskyy V, Balabanova Y, Kovalyov A, et al. (2013) Diagnostic accuracy of the genotype MTBDRsl assay for rapid diagnosis of extensively drug-resistant tuberculosis in HIV-coinfected patients. J Clin Microbiol 51: 243-248.

196. Lacoma A, Garcia-Sierra N, Prat C, Maldonado J, Ruiz-Manzano J, et al. (2012) GenoType MTBDRsl for molecular detection of second-line-drug and ethambutol resistance in Mycobacterium tuberculosis strains and clinical samples. J Clin Microbiol 50: 30-36.

197. Orikiriza P, Tibenderana B, Siedner MJ, Mueller Y, Byarugaba F, et al. (2015) Low Resistance to First and Second Line Anti-Tuberculosis Drugs among Treatment Naive Pulmonary Tuberculosis Patients in Southwestern Uganda. PLoS One 10: e0118191.

198. Singhal R, Myneedu VP, Arora J, Singh N, Bhalla M, et al. (2015) Early detection of multi-drug resistance and common mutations in Mycobacterium tuberculosis isolates from Delhi using GenoType MTBDRplus assay. Indian J Med Microbiol 33 Suppl: S46-52.

199. Finken M, Kirschner P, Meier A, Wrede A, Bottger EC (1993) Molecular basis of streptomycin resistance in Mycobacterium tuberculosis: alterations of the ribosomal protein S12 gene and point mutations within a functional 16S ribosomal RNA pseudoknot. Mol Microbiol 9: 1239-1246.

200. Okamoto S, Tamaru A, Nakajima C, Nishimura K, Tanaka Y, et al. (2007) Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria. Mol Microbiol 63: 1096-1106.

201. World Health Organization (2015). Tuberculosis country profiles: Mali. http://www.who.int/tb/country/data/profiles/en/

202. Tang X, Deng W, Xie J (2012) Novel insights into Mycobacterium antigen Ag85 biology and implications in countermeasures for M. tuberculosis. Crit Rev Eukaryot Gene Expr 22: 179-187.

203. Lavollay M, Arthur M, Fourgeaud M, Dubost L, Marie A, et al. (2008) The peptidoglycan of stationary-phase Mycobacterium tuberculosis predominantly contains cross-links generated by L,D-transpeptidation. J Bacteriol 190: 4360-4366.

204. Schoonmaker MK, Bishai WR, Lamichhane G (2014) Nonclassical transpeptidases of Mycobacterium tuberculosis alter cell size, morphology, the cytosolic matrix, protein localization, virulence, and resistance to beta-lactams. J Bacteriol 196: 1394-1402.

205. Gioffre A, Infante E, Aguilar D, Santangelo MP, Klepp L, et al. (2005) Mutation in mce operons attenuates Mycobacterium tuberculosis virulence. Microbes Infect 7: 325-334.

206. Kumar A, Bose M, Brahmachari V (2003) Analysis of expression profile of mammalian cell entry (mce) operons of Mycobacterium tuberculosis. Infect Immun 71: 6083-6087.

207. McGuire AM, Weiner B, Park ST, Wapinski I, Raman S, et al. (2012) Comparative analysis of Mycobacterium and related Actinomycetes yields insight into the evolution of Mycobacterium tuberculosis pathogenesis. BMC Genomics 13: 120.

208. Gopinath K, Moosa A, Mizrahi V, Warner DF (2013) Vitamin B(12) metabolism in Mycobacterium tuberculosis. Future Microbiol 8: 1405-1418.

209. Agarwal N, Lamichhane G, Gupta R, Nolan S, Bishai WR (2009) Cyclic AMP intoxication of macrophages by a Mycobacterium tuberculosis adenylate cyclase. Nature 460: 98-102.

210. Lu P, Lill H, Bald D (2014) ATP synthase in mycobacteria: special features and implications for a function as drug target. Biochim Biophys Acta 1837: 1208-1218.

211. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, et al. (2005) A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. Science 307: 223-227.

212. Gehre F, Antonio M, Otu JK, Sallah N, Secka O, et al. (2013) Immunogenic Mycobacterium africanum strains associated with ongoing transmission in The Gambia. Emerg Infect Dis 19: 1598-1604.

213. Niemann S, Rusch-Gerdes S, Joloba ML, Whalen CC, Guwatudde D, et al. (2002) Mycobacterium africanum subtype II is associated with two distinct genotypes and is a major cause of human tuberculosis in Kampala, Uganda. J Clin Microbiol 40: 3398-3405.

214. Simeone R, Bottai D, Brosch R (2009) ESX/type VII secretion systems and their role in host-pathogen interaction. Curr Opin Microbiol 12: 4-10.

215. Houben EN, Korotkov KV, Bitter W (2014) Take five - Type VII secretion systems of Mycobacteria. Biochim Biophys Acta 1843: 1707-1716.

216. Pym AS, Brodin P, Brosch R, Huerre M, Cole ST (2002) Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines Mycobacterium bovis BCG and Mycobacterium microti. Mol Microbiol 46: 709-717.

217. Pym AS, Brodin P, Majlessi L, Brosch R, Demangel C, et al. (2003) Recombinant BCG exporting ESAT-6 confers enhanced protection against tuberculosis. Nat Med 9: 533-539.

218. Williams M, Mizrahi V, Kana BD (2014) Molybdenum cofactor: a key component of Mycobacterium tuberculosis pathogenesis? Crit Rev Microbiol 40: 18-29.

219. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS (2003) Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. J Biol Chem 278: 41148-41159.

220. Gopinath K, Venclovas C, Ioerger TR, Sacchettini JC, McKinney JD, et al. (2013) A vitamin B(1)(2) transporter in Mycobacterium tuberculosis. Open Biol 3: 120175.

221. Shenoy AR, Sivakumar K, Krupa A, Srinivasan N, Visweswariah SS (2004) A survey of nucleotide cyclases in actinobacteria: unique domain organization and expansion of the class III cyclase family in Mycobacterium tuberculosis. Comp Funct Genomics 5: 17-38.

222. Banu S, Honore N, Saint-Joanis B, Philpott D, Prevost MC, et al. (2002) Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens? Mol Microbiol 44: 9-19.

223. Thorel MF (1980) Isolation of Mycobacterium africanum from monkeys. Tubercle 61: 101-104.

224. Thorel MF (1980) [Mycobacteria identified in a centre for veterinary research between 1973 and 1979 (author's transl)]. Ann Microbiol (Paris) 131: 61-69.

225. Coscolla M, Lewin A, Metzger S, Maetz-Rennsing K, Calvignac-Spencer S, et al. (2013) Novel Mycobacterium tuberculosis complex isolate from a wild chimpanzee. Emerg Infect Dis 19: 969-976.

226. Rahim Z, Mollers M, te Koppele-Vije A, de Beer J, Zaman K, et al. (2007) Characterization of Mycobacterium africanum subtype I among cows in a dairy farm in Bangladesh using spoligotyping. Southeast Asian J Trop Med Public Health 38: 706-713.

227. Alfredsen S, Saxegaard F (1992) An outbreak of tuberculosis in pigs and cattle caused by Mycobacterium africanum. Vet Rec 131: 51-53.

228. Gudan A, Artukovic B, Cvetnic Z, Spicic S, Beck A, et al. (2008) Disseminated tuberculosis in hyrax (Procavia capensis) caused by Mycobacterium africanum. J Zoo Wildl Med 39: 386-391.

Curriculum Vitae

Kathryn Winglee **March 24, 2015**
11 N Washington St, Baltimore, MD 21231
kwinglee@jhmi.edu
206-850-8540

Education

- Johns Hopkins University, Baltimore, MD: August 2009-present
  - Expected degree: Ph.D. in Cellular and Molecular Medicine 2015
- University of Washington, Seattle, WA: September 2005-June 2009
  - Graduated magna cum laude
  - Bachelor of Science in Computer Science
  - Bachelor of Science in Microbiology with Distinction

Research Experience

- Center for Tuberculosis Research, Department of Medicine, Division of Infectious Diseases, Johns Hopkins University (August 2009 -present)
  PhD. Thesis Research
  Advisor: Dr. William Bishai

  Projects:
  1. Examined the interaction between the microbiota and murine host during *Mycobacterium tuberculosis* infection
     - Skills:
       - Mouse models of *Mycobacterium tuberculosis* infection
       - RNA and DNA isolation from stool samples
       - Analyzed 16S, metagenomic and metatranscriptomic data
  2. Identified the mechanism of resistance to a novel anti-mycobacterial compound and determined of the function of an unannotated protein
     - Skills:
       - Molecular biology techniques, including cloning and real-time PCR
       - Ran and maintained the Ion Torrent Personal Genome Machine (PGM)
       - Analyzed next generation sequencing results from PGM (whole genome sequencing and RNA-seq), including short read alignment, SNP and indel calls, and differential expression analysis
  3. Performed whole genome sequencing of clinical isolates from Mali
     - Skills:
       - Analyzed sequencing results from Illumina, SOLiD, and PacBio
       - Cultured and maintained *Mycobacterium tuberculosis* and *Mycobacterium africanum* clinical isolates, including drug-resistant strains, under BSL3 conditions
  4. Developed software to predict the operons of *Mycobacterium tuberculosis* from RNA-seq data
     - Skills:

        - Analyzed RNA-seq data from SOLiD
        - Developed and validated user-friendly standalone Java software
    5. Determined the role of the innate immune system antimicrobial peptide cathelicidin in *Mycobacterium tuberculosis* infection
        - Skills:
            - Cultured murine and human cell lines
            - Isolation of murine bone-marrow derived macrophages and dendritic cells
            - FACS
            - Protein purification and detection, including Western blots and ELISA
            - Administration of compounds to mice via oral gavage
            - Mouse breeding

- Department of Microbiology, University of Washington (October 2007-June 2009)
  Undergraduate research
  Advisor: Professor Lalita Ramakrishnan

  Projects:
  1. Created a program, fluorescence pixel count (FPC), to quantify zebrafish bacterial burden as a high-throughput method to replace colony forming units (CFU) counts
     - Skills: zebrafish husbandry and infection
  2. Developed a program to track fluorescent cells in 3D confocal timelapses to study changes in zebrafish immune cell morphology and motility as a result of *Mycobacterium marinum* infection
     - Skills: fluorescence microscopy

- Genomation Lab, Department of Electrical Engineering, University of Washington (June 2005-August 2006)
  Undergraduate research
  Advisors: Professors Mark Holl and Deirdre Meldrum

  Project: Designed, fabricated, and tested an array of 2x2 mm heaters as well as a microfluidic module to perform Linear-After-The-Exponential Polymerase Chain Reaction (LATE-PCR) on single cells
     - Skills: microfabrication techniques

Professional Development

- Visiting graduate student at the Broad Institute with Ashlee Earl and Harvard School of Public Health with Curtis Huttenhower (June-August 2014)
- Wellcome Trust Open Door Workshop: Working with Pathogen Genomes (November 2011)
- Teaching assistant for Dr. Kendall Gray for Microbiology 302 (General Microbiology Laboratory for non-majors) at the University of Washington (September-December 2008)

- Washington NASA Summer Undergraduate Research Program (June-August 2005-2006 and 2008)

Publications

Published manuscripts:

1. S Lun, D Miranda, A Kubler, H Guo, MC Maiga, **K Winglee**, S Pelly, WR Bishai. Synthetic lethality reveals mechanisms of *Mycobacterium tuberculosis* resistance to β-lactams. *mBio* 2014; 5(5):e01767-14.
2. A Kubler, B Luna, C Larsson, NC Ammerman, BB Andrade, M Orandle, K Bock, Z Xu, U Bagci, D Mollura, J Marshall, J Burns, **K Winglee**, B Ahmadou Ahidjo, L Cheung, M Klunk, S Jain, NP Kumar S Babu, A Sher, JS Friedland, PTG Elkington, WR Bishai. *Mycobacterium tuberculosis* dysregulates MMP/TIMP balance to drive rapid caviation and unrestrained bacterial proliferation. *Journal of Pathology* 2014
3. **K Winglee**, E Eloe-Fadrosh, S Gupta, H Guo, C Fraser, W Bishai. Aerosol *Mycobacterium tuberculosis* infection causes rapid loss of diversity in gut microbiota. *PLoS ONE* 2014; 9(5):e97048.
4. S Gupta, KA Cohen, **K Winglee**, M Maiga, B Diarra, WR Bishai. Efflux inhibition with verapamil potentiates bedaquiline in *Mycobacterium tuberculosis*. *Antimicrobial Agents and Chemotherapy* 2014; 58(1):574-6.
5. K Takaki, JM Davis, **K Winglee**, L Ramakrishnan. Evaluation of the pathogenesis and treatment of *Mycobacterium marinum* infection in zebrafish. *Nature Protocols* 2013; 8(6):1114-24.
6. M Maiga, S Lun, H Guo, **K Winglee**, NC Ammerman, WR Bishai. Risk of tuberculosis reactivation with tofacitinib (CP-690550). *Journal of Infectious Diseases* 2012; 205(11):1705-8.
7. KN Adams, K Takaki, LE Connolly, H Wiedenhoft, **K Winglee**, O Humbert, PH Edelstein, CL Cosma, L Ramakrishnan. Drug tolerance in replicating mycobacteria mediated by a macrophage-induced efflux mechanism. *Cell* 2011; 145(1):39-53.

Manuscripts under review:

8. A Kubler, C Larsson, B Luna, B Andrade, EP Amaral, M Orandle, K Bock, N Ammerman, M Urbanowski, L Cheung, **K Winglee**, M Halushka, J Park, A Sher, J Friedland, P Elkington, W Bishai. The Collagenase Cathepsin K is Associated with Cavitation and Collagen Turnover in Pulmonary Tuberculosis. Submitted, *PNAS.*
9. S Pelly, **K Winglee**, WR Bishai, G Lamichhane. REMap: Operon Map of *M. tuberculosis* based on RNA Sequence Data. Submitted, *Tuberculosis*.

Manuscripts in preparation:

10. **K Winglee**, S Gupta, G Abu-Ali, C Huttenhower, A Earl, WR Bishai. The innate immune system mediates rapid changes in the gut microbiota in response to *Mycobacterium tuberculosis* infection.
11. **K Winglee**, AM McGuire, M Maiga, T Abeel, T Shea, CA Desjardins, B Diarra, B Baya, M Sanogo, S Diallo, AM Earl, WR Bishai. Whole genome sequencing of *Mycobacterium africanum* strains from Mali provides insights into the mechanisms of geographic restriction.
12. **K Winglee**, S Lun, M Pieroni, A Kozikowski, WR Bishai. Mutation of *Rv2887*, a marR-like gene, confers *Mycobacterium tuberculosis* resistance to a pyrido [1,2-*a*]benzimidazole-based agent.
13. S Gupta, **K Winglee**, R Gallo, W Bishai. Cathelicidins link innate and adaptive immune responses against *Mycobacterium tuberculosis* in the mouse model of infection.

Oral Presentations

- Genomic evaluation of M.TB drug targets by deep sequencing. Johns Hopkins Tuberculosis Day, June 9, 2011.
- A Tool for Rapidly Identifying Strain Differences in Deep Sequencing Data. Joint Keystone Symposia on Tuberculosis: Immunology, Cell Biology and Novel Vaccination Strategies (J3) and Mycobacteria: Physiology, Metabolism and Pathogenesis – Back to the Basics (J4), January 18, 2011.
- Allelic exchange, clinical isolates, and deep sequencing with *Mycobacterium tuberculosis*: a cautionary tale. K-RITH Research in Progress, September 29, 2010.

Posters

1. **K Winglee**, S Lun, W Bishai. Rv2887 is a Transcriptional Regulator and Potential Drug Target. Keystone Symposia on Novel Therapeutic Approaches to Tuberculosis (C7), April 2014.
2. S Gupta*, K Cohen, S Tyagi, **K Winglee**, M Maiga, B Diarra, W Bishai. Increased anti-mycobacterial activity of bedaquiline with the efflux pump inhibitor verapamil. Keystone Symposia on Novel Therapeutic Approaches to Tuberculosis (C7), April 2014. *presenting author
3. **K Winglee**, E Eloe, S Gupta, H Guo, S Lun, C Fraser, W Bishai. Changes in the Gut Microbiota of Mice with *Mycobacterium tuberculosis* Infection. Keystone Symposia on Host Response in Tuberculosis (X7), March 2013.

4. S Gupta*, **K Winglee**, R Gallo, W Bishai. Cathelicidins: An Important Link Between Innate And Adaptive Immune Responses Against *Mycobacterium tuberculosis*. Keystone Symposia on Host Response in Tuberculosis (X7), March 2013. *presenting author
5. **K Winglee**, S Lun, D Geiman, M Pieroni, A Kozikowski, W Bishai. A Tool for Rapidly Identifying Strain Differences in Deep Sequencing Data. Keystone Symposia on Tuberculosis: Immunology, Cell Biology and Novel Vaccination Strategies (J3), January 2011.
6. CT Yang*, JM Davis, **K Winglee**, CJ Cambier, C Hall, P Crosier, L Ramakrishnan. Cellular dynamics of neutrophils during early mycobacterial infection. Keystone Symposia on Tuberculosis: Immunology, Cell Biology and Novel Vaccination Strategies (J3), January 2011. *presenting author
7. S Lun*, D Geiman, H Guo, **K Winglee**, R Morris, W Bishai, CJ Thompson. Identification of a massive genomic duplication in a clinical isolate of *Mycobacterium tuberculosis* of the Euro-American lineage by deep sequencing. Keystone Symposia on Mycobacteria: Physiology, Metabolism and Pathogenesis – Back to the Basics (J4), January 2011. *presenting author

---

Professional affiliations

- AAAS/Science
- American Society for Microbiology
- Washington NASA Space Grant alumni

Scholarships/Awards/Fellowships

- William and Mary Drescher Award (2009-2010)
- University of Washington Dean's List (2005-2009)
- Levinson Emerging Scholars Program (2008-2009)
- Wisniewski Endowed Scholarship (2008)
- Mary Gates Research Scholarship (2008)
- Washington NASA Space Grant Scholarship (2005-2008)
- University of Washington Undergraduate Scholar Award (2005-2006)